Parameter Tuning for Unit Selection Speech Synthesis

Joanna Keating

Supervisors: Dr. Rob Clark & Dr. Cassie Mayo

Master of Science

in

Speech and Language Processing

Theoretical and Applied Linguistics

School of Philosophy, Psychology and Language Sciences

University of Edinburgh

2005

Abstract

This project aims to contribute to current research on the quality of speech synthesis by conducting a perceptual experiment to discover a better set of target cost weights for the Festival speech synthesis system. From the experiment, the acoustic parameters that listeners use when judging synthetic speech will become clearer, as will the importance that each parameter has.

The project uses unit selection synthesis, which chooses units for concatenation using a series of target and join costs. Each cost is assigned a weight value which indicates its importance in the overall cost. This project manipulates the target cost weight values in order to find a set of values that better represents the listeners' perception of the quality of the synthetic speech.

Previous research shows that perceptual experiments are a common way of evaluating the quality of speech synthesis, and this project uses a listening experiment consisting of paired comparisons to reveal information about how listeners judge synthetic speech. The results from the experiment were analysed using multidimensional scaling to show the structure of the data and provide insight into the processes involved in speech perception.

The results showed that when judging synthetic speech, participants pay attention to position in phrase, position in syllable, and stress parameters. It was also found that participants grouped the stimuli on the basis of which of these parameters was given the weight value of 1. The results also showed that a lack of weight on these parameters has more effect on the selection of units from the database than a large amount of weight. Through analysis of the results it was shown that position in syllable was the most important parameter for high quality speech.

Acknowledgements

I would like to thank my supervisors, Dr. Rob Clark and Dr. Cassie Mayo, for their valuable ideas and suggestions throughout this project. Their guidance and support was very helpful and much appreciated. I would also like to thank Jeff Mitchell, who spent countless hours throughout this MSc patiently explaining such simple things to me as computer programming. Very special thanks go to Evia Kainada and Fiona Skilling, for their proofreading and suggestions on the draft of this report. Many thanks also to all the people who took part in my experiment, whose time, effort, and participation were an essential part of this project.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified.

(*Joanna Keating*)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

The evaluation of speech synthesis is a very complicated and important issue, providing valuable information which can be used to improve the quality of speech in the growing number of speech synthesis applications. This view is supported by ISCA SynSIG (International Speech Communication Association Speech synthesis Special Interest Group), who in 2003 listed the evaluation of synthetic speech as one of the top issues in speech synthesis that requires more research (King et al., 2005). Therefore, this project aims to evaluate synthetic speech from the Festival speech synthesis program developed at Edinburgh University, making it a timely and appropriate step in the direction of speech synthesis quality research. The goal of this project is to determine a better set of target cost weight values which will allow the Festival program to produce more natural-sounding synthesis. The project aims to uncover important information about the acoustic parameters that listeners use when judging the quality of synthetic speech, and to manipulate these parameters to discover the optimal weight values for them. The results from this project could then be implemented in the Festival program in order to improve it.

The project consists of a perceptual experiment which aims to reveal more information about the relationship between target costs and the perceived discontinuity of synthetic speech. Although computational measures have also been used to evaluate synthesis systems, they are poor at evaluating the perceived quality of synthetic speech (King et al., 2005). This is because the quality of speech is a psycho-physical construct (King et al., 2005), and can not accurately be measured objectively or physically - it is instead an interaction between an acoustic voice stimulus and a listener (Kreiman and Gerratt, 1998). Therefore, most studies relating target and join costs with auditory judgements of discontinuities have focused on using perceptual data (Syrdal and Conkie, 2004). Even though perceptual evaluation of speech synthesis systems is often costly and time-consuming (King et al., 2005), synthesis systems using weights determined from perceptual experiments consistently produce smoother transitions and better voice quality than systems using weights that have been set manually (Lee et al., 2001). For this reason this project uses perceptual data achieved through a listening experiment to determine better values for Festival's target cost weights.

Determining the weights of cost terms in unit selection is a very difficult task, as the relationship between the properties of two concatenated units and their perceived quality is an area that is not well understood (Lee et al., 2001; Stöber et al., 2001; Syrdal and Conkie, 2004). Therefore this project attempts to uncover the acoustic parameters which are important to the listener when making a judgement regarding the quality of synthetic speech. It is hoped that once identified, the weights for these parameters can be manipulated and a perceptual experiment will show the optimal values for these weights. In this way the project aims to gain a better understanding of how target costs relate to the perceived quality of synthetic speech.

As Allen and Scollie (2002) state, "the manner in which listeners weight acoustic information can be studied using paired comparisons procedures and multidimensional scaling". For this reason a perceptual experiment involving paired comparisons was conducted, and the responses were analysed using

multidimensional scaling (MDS). MDS techniques have been used to evaluate participants' perceptions of a variety of complex stimuli and have the potential to provide subtle insights into the processes involved in speech perception (Barry et al., 2002; Allen and Scollie, 2002). It has also been shown that MDS is able to determine what and how many acoustic features are perceptually important to a listener (Barry et al., 2002).

This chapter contains an introduction to unit selection and the Festival speech synthesis program, which is used in this project. Target costs are also introduced, followed by previous work concerning target and join costs. Chapter 2 describes the design and procedure of the experiment, as well as the preliminary work done regarding the manipulation of the target cost weights. An introduction to multidimensional scaling is also included, followed by previous work done regarding MDS and perceptual experiments in the linguistics field. Chapter 3 includes the results from the experiment and an analysis and discussion of these results.

1.2 Unit Selection

In unit selection speech synthesis, synthetic speech is produced by concatenating small units of real speech. First a large database of natural speech is recorded by a speaker and segmented into the desired units. The most common units used in unit selection are half phones, phones, or diphones, although smaller or larger units such as triphones are possible (Clark et al., 2004; Lee et al., 2001). The units in the database are labelled using dynamic time warping, which aligns the recorded speech with synthesised speech and compares the waveforms to identify the phone boundaries. Once the database has been segmented and labelled it contains many examples of the same unit with varied prosodic and spectral properties. The labelled units are then the candidate units from which the best ones will be chosen during synthesis. Because unit selection is able to specifically choose the best units for synthesis, it has

the potential for higher quality and more natural-sounding speech (Syrdal and Conkie, 2004). However, it requires an algorithm to select the most appropriate units available to construct the utterance at run-time. Unit selection speech quality also depends on corpus size and characteristics (Adachi et al., 2005), as a large corpus will have more examples of each unit, and will therefore be able to produce better quality synthesis.

Unit selection synthesis differs from standard diphone synthesis because a unit selection database contains many examples of the same unit. A diphone synthesis database only contains one example of each diphone, which means that it does not have any choice when choosing diphones for synthesis. Diphone synthesis therefore has to predict many properties of each diphone, such as duration and prosody, in order to produce synthetic speech that contains such characteristics which are consistent throughout the utterance. An advantage of unit selection synthesis is that much of this information does not have to be predicted (Clark et al., 2004). Because the natural durations and prosody from each unit in the database are used, these properties do not need to be modelled by the synthesis system. As signal processing (such as PSOLA) can cause the synthesis quality to deteriorate, using the natural characteristics of a unit allows for better quality synthesis (Hunt and Black, 1996; Adachi et al., 2005). Therefore, because a unit selection database contains a large number of units with varied prosodic and spectral characteristics, unit selection is able to produce more natural-sounding speech than diphone synthesis (Lee et al., 2001; Hunt and Black, 1996).

Unit selection works by taking a string of text as input and performing a series of linguistic analyses on it. These steps produce a target specification, which defines the desired units needed to synthesise the input text. The target specification also includes prosodic features which specify characteristics of the target speech, such as duration and pitch. The database is then searched for all appropriate units needed to make up the target utterance. From these proposed units, the optimal units are chosen using a Viterbi search. The optimal units are those which most closely match the target specification, thereby

minimising the amount of signal processing required, which minimises the distortion of the natural waveform of the unit (Hunt and Black, 1996).

The Viterbi search works by selecting the sequence of units with the minimum overall cost. The minimum overall cost is a weighted sum of two costs: the *target cost* and the *join cost* (also known as the concatenation cost). The target cost indicates "how closely a candidate unit in the database matches the specification of the target unit", while the join cost indicates "how well two adjacent units can be joined" (Vepa and King, 2004b). Therefore, the minimum overall cost indicates the sequence of units that best matches the target specification, and therefore produces the best quality synthesis (Syrdal and Conkie, 2004). In other words, the minimum overall cost attempts to predict which concatenations will sound natural to the listener, and which will sound unnatural. It is therefore important that the costs reflect what listeners hear as good and bad synthesis.

In order to make sure that a low overall cost sounds more natural to the listener and a high overall cost sounds more unnatural, each target cost is assigned a weight value which represents its importance in the selection of units from the database. By assigning appropriate weights to each cost, the relationship between the cost and the listener's perception can be modelled more accurately. This project attempts to improve the current target cost weights in the Festival speech synthesis program in order to allow the system to produce better quality synthetic speech.

1.3 Festival

This project was carried out using the Festival Speech Synthesis System, which was developed by the Centre for Speech Technology Research (CSTR) at Edinburgh University. Festival is a speech synthesis program that is capable of text-to-speech synthesis using both diphone synthesis and unit selection synthesis. The synthesis process is "like a chain, whereby an utterance structure is built

in stages" which modify the text and produce the final output as a synthesised speech waveform (Black et al., 2000). Previously, the available unit selection synthesis was only suitable for limited domains and used cluster unit synthesis (Clark et al., 2004). However, this form of synthesis is no longer the most current (Clark et al., 2004), and this project uses a version of Festival which contains a general purpose unit selection engine called *multisyn*.

1.3.1 Multisyn

Multisyn was designed to address the previous problems in Festival's synthesis techniques in order to produce a better system which is more robust and versatile (Clark et al., 2004). The multisyn engine is therefore better able to deal with text from multiple sources and is not confined to limited domains. It can also be used as a development tool for further research of speech synthesis and natural language generation, and the voice building process is quite straightforward (Clark et al., 2004). This has been done so that even people with a limited amount of knowledge about voice building can easily build new voices.

The multisyn unit selection engine uses diphones and works using three steps. First, linguistic analysis is performed on the text which generates a target utterance. For this target utterance, suitable candidate units are then selected from the database and proposed as possible units for synthesis. The best sequence of candidate units is then found using the Viterbi algorithm, which chooses the sequence with the lowest overall cost. The voice used in the synthesis is a CSTR prototype voice called *cstr-nina*, which has almost 175,000 phones taken from 5.8 hours of speech (Clark et al., 2004). The overall cost is again made up of target and join costs, the current weightings for which have been derived empirically to provide a baseline acceptable performance (Clark et al., 2004). However, little previous research has been conducted to improve these costs, and as Clark et al. (2004) state, these weights can easily be changed to values based on perceptual evaluation. This project aims to contribute to the

improvement of Festival's multisyn engine by providing experimental results to indicate better values for the weights of the target costs.

1.3.2 Target Costs

When the unit selection database is created, standard signal processing techniques are used to obtain identical information about each unit in the database, characterising each one by a multidimensional feature vector (Hunt and Black, 1996). As a result of the signal processing done to the input text at synthesis time, the target units are also characterised with feature vectors containing the same information as the feature vectors for each unit in the database. The target cost is then calculated as the weighted sum of the differences between the elements of the target and candidate feature vectors (Hunt and Black, 1996). In this way the target cost represents how similar a candidate unit is to the target unit.

While many other systems use anywhere from four to thirty cost terms (Prudon and d'Alessandro, 2001), Festival's multisyn engine uses only eight target costs and three join costs. The three join costs are for pitch, energy, and spectral mismatch, but as this project focuses on target costs, these were excluded from the script when the experimental processes were carried out. The current target costs and weights are shown in Table 1.1.

1.4 Previous Work

Previous research has mainly focused on the improvement of join costs, while target costs have not received much attention. As there was a limited amount of previous work focusing on target costs, research regarding join costs has also been summarised here, as the two topics are very much related.

Position in Phrase	20.0
Stress	10.0
Punctuation	10.0
Part of Speech	6.0
Position in Syllable	5.0
Position in Word	5.0
Left Context	4.0
Right Context	3.0

Table 1.1: Current target cost weights

In their article, Syrdal and Conkie (2004) introduce new data-driven methods of determining perceptually-based join costs which use human perceptual data to train join cost models. First a perceptual study was done which recorded the detectability of concatenation discontinuities. This observed probability of listeners detecting a discontinuity for a given join was then used to train and test the join cost models, which predicted perceptually-based join costs. Both linear regression and Classification and Regression Tree (CART) models were used, and each was trained on several different sets of predictor variables. The results showed that the two models achieved very similar results in predicting human detection rates.

As most conventional costs ignore the original prosody of speech segments, Adachi et al. (2005) proposed a target cost for prosody to evaluate prosodic differences between target prosody and speech candidates. Since speech quality degrades in proportion to the amount of prosody modification (by analysis-synthesis or the PSOLA method), it is important to choose units from the database which match the target as closely as possible in terms of prosody so that little or no modification is needed. Nine databases of Japanese speech with different prosodic characteristics were recorded and the relationships between the amount of prosodic modification and the perceptual degradation

was investigated. The results showed that the amount of perceptual degradation differs according to the prosodic features of the original speech, and so a new cost function which changes according to the prosody of a speech database was proposed. The new target cost consists of two sub-costs: the first is the difference between target F0 and duration and those of the candidate unit, and the second is the phonemic difference between the target unit and the candidate unit. When a perceptual test of quality was conducted, the results showed that the proposed cost function generated higher quality speech than the conventional cost functions.

In their article, Hunt and Black (1996) propose a new way of looking at unit selection where the units in the database can be considered as states in a state transition network. Because any unit can potentially be followed by any other, the network is fully connected. This means that in order to synthesise a waveform, the path through the state transition network (the sequence of database units) with the minimum overall cost must be found. The state occupancy cost is given by the target cost and the state transition cost is given by the join cost. Given these two costs, the best units for synthesis are then chosen using a pruned Viterbi search. The paper provides two methods for training weights for the cost functions from natural speech - weight space search and regression training. The results showed that both methods provided weights which produced more natural speech than can be obtained by hand-tuning.

Lee et al. (2001) propose a new method for setting weights in the cost functions which is based on a perceptual preference test. The new method uses an algorithm that searches for a set of weights that can produce a ranking of utterances that is close to the ranking achieved from perceptual test results. First, listeners were asked to rank a set of different synthesised versions of words in order of their preference for each version. Using a heuristically chosen set of initial weights, the total cost of each version was calculated, and the versions were ranked according to these costs. The ranking was then compared to the ranking obtained from the perceptual test. The weights were then adjusted, the total cost re-calculated, and the versions re-ranked in an attempt to

achieve a closer match to the ranking achieved from the perceptual test. The results showed that this approach can successfully predict human preference patterns, and that in 83% of the cases, the algorithm using the optimal set of weights chose the same version that human listeners preferred.

In their paper, Stöber et al. (2001) presented an objective distance measure which was used to sort the units in a unit selection database in relation to a given natural unit. The units were sorted according to the computed similarity between the natural unit and the database unit. Then a perceptual experiment was done which presented participants with a set of three synthesised stimuli. Participants then had to compare the set of stimuli to the natural version and choose which of the three was most similar to the natural version. The results showed that the listeners' preferences reflected the ranking chosen by the objective distance measure. However, the more the listeners disagreed with the algorithm, the more they also disagreed with each other, suggesting that if stimuli are too similar, they become very difficult to classify.

Prudon and d'Alessandro (2001) describe the development of a text-to-speech synthesis system in French that uses only four selection criteria. The system first converts the input text into a phoneme chain containing prosodic and accentual descriptions. Then the optimal sequence of segments is chosen using a dynamic programming algorithm. The join cost is made up of two criteria - one to select adjacent phones in the same speech segment where possible, and one to prevent large pitch differences between adjacent phones. The target cost is also made up of two criteria - one to find a phone with the best position in the corresponding phonetic word, and one to find a phone with the best position in the corresponding phonetic syllable. The weights for the two target costs were chosen by a simple trial-and-error procedure which assigned 50% of the total target cost weight to each one. A perceptual experiment was then conducted which presented synthesised sentences or phrases. Participants were required to rank each stimulus according to seven criteria such as voice pleasantness and pronunciation. The results showed that the new system was preferred to the previous system along all seven evaluation categories.

In their paper, Vepa and King (2004a) present their research which makes objective comparisons between different join cost functions. This was done using only a single set of perceptual tests. Synthesised utterances were presented to the participants, who were required to rate the quality of the join presented in the utterance. The stimulus set contained utterances with a wide range of qualities of join and was designed to measure the degree of perceived concatenation discontinuity. Join cost functions were then designed to maximise the correlation with the perceptual ratings. The join cost functions were then compared by computing the correlations between the perceptual ratings and the various join costs. The different join costs were based on three speech parameterisations (MFCCs, LSFs, and MCA coefficients) and four distance measures (Euclidean, absolute, Mahalanobis, and symmetric KL). The results showed that Mahalanobis distance using LSFs plus delta coefficients was an appropriate join cost, as was KL distance using MCAs.

In another article, Vepa and King (2004b) continued their previous work, which subjectively evaluated join cost functions derived from spectral distances. The join costs were proposed in an earlier paper, and this work evaluated the top three join cost functions revealed in that study. Each of the three join cost functions was combined with three different smoothing methods, and the different combinations were evaluated in a perceptual test. The results showed an agreement with the rankings obtained in the previous experiment.

1.5 Summary

This project aims to contribute to current research on the quality of speech synthesis by conducting a perceptual experiment to discover a better set of target cost weights for the Festival speech synthesis system. From the experiment, the acoustic parameters that listeners use when judging synthetic speech will become clearer, as will the importance that each parameter has.

The project uses unit selection synthesis, which chooses units for concatenation using a series of target and join costs. Each cost is assigned a weight value which indicates its importance in the overall cost. This project manipulates the target cost weight values in order to find a set of values that better represents the listeners' perception of the quality of the synthetic speech.

Previous research shows that perceptual experiments are a common way of evaluating the quality of speech synthesis. Therefore, this project uses a listening experiment consisting of paired comparisons to reveal information about how listeners judge synthetic speech. This information is then used to determine a better set of target cost weight values for Festival.

Chapter 2

Experiment

2.1 Introduction

This chapter discusses the experiment carried out during the project. Before the experiment could be run, preparation work involving parameter manipulations had to be done to create the stimuli. A preliminary analysis of the initial stimuli was then carried out to produce a ranked order of parameter combinations, and is discussed below. This chapter also details the design of the experiment, including the materials, participants, and procedure used. An introduction to multidimensional scaling is then given, followed by a summary of previous work using multidimensional scaling for speech perception experiments. A short explanation of the analysis used on the data is also given.

2.2 Preparation

2.2.1 Parameter Manipulation

The first step was to experiment with the values of the target cost weights to find any preliminary conclusions which could be used to guide the remainder

of the project. Seventeen sentences from the TIMIT corpus were used as test sentences, which were then synthesised using different parameter combinations. The first trial was synthesised using the original parameters as listed in the previous chapter in Table 1.1, with the inclusion of two join cost parameters: F0 and duration. This combination was the original parameter combination that was in use before the start of this project. The second trial used the original values for the target costs, but changed the two join cost parameters to 0, as join cost was not the focus of this project and therefore would not be included in any more of the trials. These first two tests were regarded as the baseline to which all future tests were compared.

Parameters were changed by altering the values for the weights directly in the script. Festival was then reloaded and the same set of 17 sentences was synthesised again with the new weights and saved in a separate file. The list of sentences used is shown at the end of this report in Appendix A.

Parameter weights are mathematically changed into a percent out of 100 when the script is compiled, therefore a value of 10 is weighted twice as heavily as a value of 5, and a value of 15 is three times as heavily weighted as a value of 5.

2.2.1.1 First Set

The first set of parameter manipulations involved changing all the values to 0 except for one, which remained at an (arbitrarily) assigned value of 5. When this value was then converted it was actually given 100% of the weight. Then one by one in turn, each value was assigned the value of 5 (or 100% of the weight) while all the others were given a value of 0. The values for this first set of trials are shown in Table 2.1.

The first set allowed the sentences to be synthesised using a set of parameters where only one parameter was given any weight. This would show the effects, if any, that each parameter could have on the synthesis.

	T3	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**
Stress	**5**	0	0	0	0	0	0	0
Position in Syllable	0	**5**	0	0	0	0	0	0
Position in Word	0	0	**5**	0	0	0	0	0
Position in Phrase	0	0	0	**5**	0	0	0	0
Part of Speech	0	0	0	0	**5**	0	0	0
Punctuation	0	0	0	0	0	**5**	0	0
Left Context	0	0	0	0	0	0	**5**	0
Right Context	0	0	0	0	0	0	0	**5**

Table 2.1: Trials 3 - 10 showing the first set of parameter manipulations

2.2.1.2 Second Set

The second set of parameter manipulations involved changing all the values to 5 except for one, which was given the value of 0. Again, each separate trial involved changing a different parameter to the value of 0, while the rest remained at 5, giving them each an equal weight. The values for the second set of trials are shown in Table 2.2.

	T11	**T12**	**T13**	**T14**	**T15**	**T16**	**T17**	**T18**
Stress	**0**	5	5	5	5	5	5	5
Position in Syllable	5	**0**	5	5	5	5	5	5
Position in Word	5	5	**0**	5	5	5	5	5
Position in Phrase	5	5	5	**0**	5	5	5	5
Part of Speech	5	5	5	5	**0**	5	5	5
Punctuation	5	5	5	5	5	**0**	5	5
Left Context	5	5	5	5	5	5	**0**	5
Right Context	5	5	5	5	5	5	5	**0**

Table 2.2: Trials 11 - 18 showing the second set of parameter manipulations

This second set of trials allowed the sentences to be synthesised using a set of parameters where all the parameters were given equal weight except for one, which was not given any weight. This would show the effects, if any, that the loss of each parameter could have on the synthesis.

2.2.2 Preliminary Analysis

The original values produced synthetic speech that was quite comprehensible and fairly natural in terms of prosody and duration. During the second trial when prosody and duration were set to 0, synthesis quality decreased and there were quite a few problems with incorrect prosodic structures and strange pitch changes.

The first set of trials showed that when Stress was given a value of 5 (all the weight), the synthesis quality was not as good as the original, however, each word was pronounced quite well, and the decrease in quality came from the absence of any natural declination at the end of the sentence. The trial which gave Stress all the weight was also the first trial which managed to correctly pronounce the word "needle" (other trials produced a variation on /nɪdəl/). When Stress was weighted heavily the synthesis also performed better on acronyms such as "KLM" and "BMI". Over the next few trials it was shown that when Position in Phrase was given all the weight, the synthesis produced much more natural-sounding declination.

Overall it was difficult to tell if there were differences between the trials, and it was unclear which set of parameters produced overall better synthesis, as some sentences came out quite clear while others were almost incomprehensible. For this reason I decided to listen to each sentence individually and rank them according to how natural each one sounded.

2.2.2.1 Ranking

All 289 utterances were separated into groups of the same sentence, making 17 groups in total. Each sentence in a group had been produced by one of the different trials, or in other words, one of the different sets of parameter weights. All versions of a single sentence were then listened to and ranked according to how natural they sounded. The ranking was done simply by listing the trial numbers in order from the most natural to the least natural. This was repeated for each sentence, and the rankings were kept separate.

Once all the trials had been ranked in order of their naturalness for each sentence, the rankings for each sentence were converted into scores by assigning the trial in first place the number 1, the second place trial the number 2, and so on. At the end of this ranking, each trial (parameter combination) had 17 different scores representing how natural each of its 17 sentences sounded. These 17 scores were then averaged and the standard deviation was calculated. The average score gave an idea of which trials consistently produced the most natural-sounding speech, and which trials consistently produced unnatural-sounding speech. The standard deviation was used to show the range of scores that a trial received, since many trials scored quite well on some sentences, and very badly on others.

One disadvantage of ranking in this manner is that it implies that there is always the same distance between the scores. It assumes, for example, that the distance between ranks 1 and 2, and the distance between ranks 2 and 3 are the same. However, the differences between the scores for one sentence may be smaller or greater than for another sentence, and even the differences between scores in the same sentence may vary. Sometimes the variations of a sentence were very close in terms of naturalness, and sometimes they were very far apart. Therefore this system was perhaps not the best way to rank the sentences, but it served as a preliminary ranking system and provided more information.

2.2.3 Preliminary Results

The preliminary results showed that the two baseline trials produced the most consistently natural-sounding speech. This was expected as the first trial included join costs, and both trials included a range of weight values allowing some values twice the importance of others. However, already knowing this information from the start made the other trials of more interest. Of the first set of trials it was shown that the top three trials were those which placed all the weight on Stress, Position in Syllable, and Position in Phrase. The second set of trials showed that by removing any weight on Position in Phrase, Position in Word, and Position in Syllable, the naturalness of the sentences decreased the most. Interestingly, this is consistent with Prudon and d'Alessandro (2001), who chose to include word position and syllable position as their two target costs. From the preliminary results it was concluded that Stress, Position in Syllable, and Position in Phrase were the three parameters most responsible for natural-sounding speech. These parameters were then chosen as the three parameters which would be manipulated in the experiment.

2.3 Running the Experiment

2.3.1 Introduction

In this experiment I was interested in determining if changes to the values of the parameter weights would result in different diphones being chosen during synthesis time, and if these changes would have an effect on the listener's perception of the resulting synthesis. In other words, by changing the weight values for the target costs, would the listener report an audible difference in the naturalness of the synthesis, or would the changes have little or no effect? For example, if stress was weighted more heavily in the synthesis of one sentence, but not in another, would the listener hear a difference in the naturalness of the two sentences? The experiment aimed to investigate which parameter changes

had the biggest effect on the phonemes chosen for synthesis, and whether or not these changes had an effect on the listener. The experiment also aimed to uncover which values of the parameters achieved any of these effects, and ultimately which values for the weights resulted in the most natural-sounding synthesis.

2.3.2 Materials

Thirty arbitrarily chosen sentences from the TIMIT corpus were synthesised with all the weight values set at 5. This gave all parameters equal weight and ensured that any synthesis effects were not caused by a single parameter having more or less weight than any other. From these 30 sentences, 5 sentences were chosen for the experiment. Care was taken to choose sentences that were of approximately equal length and synthesis quality in order to be sure that any effects found in the experiment were truly a result of a manipulation, and not a difference which existed before the experiment began. Therefore, all sentences were between 14 and 16 syllables long. Although almost all previous experiments concerning join costs have studied audible discontinuities in vowel joins (Syrdal and Conkie, 2004; Grabe et al., 2003), this project used sentences, as it focused on target costs. Therefore full sentences were needed in order to demonstrate any possible discontinuities concerning all of the eight target costs as shown in Chapter 1. The sentences used in the experiment are listed below.

1. "Will you please describe the idiotic predicament."
2. "Only the most accomplished artists obtain popularity."
3. "Young people participate in athletic activities."
4. "Etiquette mandates compliance with existing regulations"
5. "Continental drift is a geological theory."

As mentioned above, three parameters were chosen to be manipulated: Stress, Position in Syllable, and Position in Phrase. Using these three parameters, 9 different versions of each sentence were synthesised for each of the 5 sentences, creating 45 utterances in total. The parameter combinations used for each sentence are listed in Table 2.3 below, where all other parameters were given the value 1. This set of parameter values allowed me to test whether or not certain parameters should be weighted twice, five times, or ten times as heavily as others, and also whether or not any two parameters should be weighted equally. By choosing this set of parameters I was able to see if greatly increasing or decreasing the weight of one of the chosen parameters would affect the quality of the resulting synthesis.

	V1	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**	**V9**
Stress	1	1	5	5	10	10	10	10	1
Position in Syllable	5	10	1	10	1	5	10	1	10
Position in Phrase	10	5	10	1	5	1	1	10	10

Table 2.3: Parameter combinations used to synthesise experiment sentences

Each of the 45 sentence versions were then paired with all other versions of the same sentence, making 36 pairs for each sentence. Sentences were paired with a version of the same sentence because the differences between versions were subtle, and grouping them in pairs of the same sentence made these differences easier to notice. It also allowed participants to concentrate on the specific differences between the versions of each sentence, and not to be distracted by the difference in words which would have resulted if sentences were paired with versions of another sentence. Each pair was also presented in forward and backward order, meaning that the sentences were presented as "A" followed by "B", and also as "B" followed by "A". This was done to ensure that any differences perceived between the two sentences in a pair were not due simply to the order in which they were presented. This increased the number of pairs for each sentence to 72, making 360 pairs in total for all 5 sentences. The wave-

forms of each sentence were then converted to a common 22050Hz sampling rate, 16-bit resolution, and one channel (Mono) using CoolEdit.

The total 360 stimulus pairs were split into two equal groups of 180 pairs each, and two identical experiments were designed to present them to participants. The experiments were created as two experiments instead of one in order to control which pairs of stimuli the participants heard. The stimuli were carefully split into two groups so that each experiment contained stimuli of all 5 sentences, but the pairs themselves (which version had been paired with which version) were different. This allowed the participants in each experiment to listen to half the total pairs once each forwards and backwards, while still giving as much variation as possible by presenting stimuli from all 5 sentences. Variation of the sentences was important for keeping participants interested and alert during the experiment. By making sure that each participant heard the same stimuli both forwards and backwards, any changes in responses could be checked to determine whether they were due to the order of presentation or to the participant. Presenting stimuli in both orders also allowed for a consistency check of the responses. Because synthetic stimuli that are too similar can lead to random judgements, consistent responses for pairs presented both ways would show that the stimuli were not too similar and that the responses were not random (Stöber et al., 2001).

The experiments were created in E-Prime, which is an application that allows the user to generate experiments and collect data from them. It also allows the user to combine the data from all participants and process it by filtering, editing, and exporting it to external statistical applications (Schneider et al., 2002). The two experiments were created using a template experiment designed by Dr. Cassie Mayo, and changes were then made to suit the needs of this experiment. The experiment was presented on one of three different computers, and stimuli were presented via headphones.

2.3.3 Participants

Each of the two experiments was given to 10 native English speakers, resulting in 20 participants in total. It has been demonstrated through the use of MDS techniques that the relative perceptual salience of pitch varies for speakers from different language backgrounds (Barry et al., 2002), therefore native English speakers were used to control for the possibility that non-native speakers perceive English differently, or have different opinions about what constitutes naturalness in English. By having 20 participants, each pair combination was presented and heard by 10 different participants. Because different people tend to have different preferences concerning the naturalness of speech (Stöber et al., 2001), it is important to use a large number of participants to allow any results found to be generalised to the whole native English-speaking population. Of the 20 participants, 6 were male and 14 were female. All participants were between the ages of 19 and 40, and were from Canada, the United States, New Zealand, and the UK.

2.3.4 Procedure

Participation in the experiment took place in one of three isolated experiment booths, each of which contained a computer and headphones. The experiment was presented on the computer screen and the stimuli were presented via headphones. Limited instructions were given to each participant on paper before the experiment began, as full instructions were given on the computer at the start of the experiment. This was to ensure that all participants received the same instructions for the experiment, and were not biased by anything the experimenter said before the start of the experiment.

The experiment began with a full set of instructions explaining that the participant should focus on the naturalness of each sentence in a pair. The participant should then make a decision about whether or not the two sentences in the pair were the same or different in terms of their naturalness. Participants should

then indicate their response by pressing either "0" or "1" on the keyboard, where 0 represented different, and 1 represented the same degree of naturalness. Some previous experiments chose to use a numbered ranking system, where participants indicated the perceived quality of an utterance using a linear scale (Hall, 2001; Allen and Scollie, 2002; Adachi et al., 2005; Prudon and d'Alessandro, 2001). However, different listeners assign numbers to the stimuli differently, and a listener often uses numbers inconsistently throughout the course of the experiment (Hall, 2001). It has also been shown that ranking is only appropriate for characteristics that can be consistently partitioned linearly by participants (King et al., 2005). Therefore, when evaluating speech quality many experiments use questions which do not use ranking, but instead employ paired comparisons (Syrdal and Conkie, 2004; Barry et al., 2002; Lee et al., 2001). As Hall (2001) states, the most sensitive way of comparing the quality of synthetic speech is by using paired comparisons among all versions. For these reasons this experiment used paired comparisons and did not use ranking.

The instructions were then followed by three example pairs of sentences - one of which was different in terms of naturalness, and two of which were the same. These example sentences were intended to familiarise the participant with synthesised speech, and also helped explain the instructions and the participant's task by giving examples of what might be presented in the experiment. The two sentence pairs which were presented as the same were in fact the same waveforms presented twice in a row. One was an example of a sentence pair where both parts were very unnatural, and the other was an example of a pair where both parts were very natural. The pair presented as different was agreed to be different by three linguistics students, and consisted of a pair of sentences where one part was very unnatural and the other was very natural. The example sentences were always presented in pairs of the same sentence, but were different sentences from the ones presented during the main body of the experiment. This was done to ensure that the responses gathered in the experiment had not been biased by anything in the instructions. The sentences used as examples are listed below in Figure 2.1.

"She had your dark suit in greasy wash-water all year."	*Different*
"Artificial intelligence is for real."	*Same-Natural*
"The altruistic dowager helped many malnourished vagrants."	*Same-Unnatural*

Figure 2.1: Example sentences used in experiment

Following the example sentences was a practice period where participants were told they would have the chance to practice listening to pairs of sentences and indicating their response. Four practice sentences were then presented to the participant, preceded by a reiteration of the instructions and response keys. Two of the practice sentences were the same in terms of naturalness, and two were different, and the pairs were presented in random order so as to control for presentation order. As before, the two sentence pairs which were presented as the same were actually the same waveforms presented twice in a row, and the two sentence pairs presented as different were agreed to be different by the same three linguistics masters students as above. Again, the practice sentences were always presented in pairs of the same sentence, but again they were different from the 5 sentences presented during the main body of the experiment. They were also different from the example sentences. This was again done to ensure that the responses given in the practice session and in the experiment were not biased by anything in the instructions. The responses from this practice session were then consulted during analysis of the experiment results in order to ensure that participants had chosen the correct answers and had therefore understood the instructions. The sentences used in the practice session are shown below in Figure 2.2.

Following the practice session was the main body of the experiment. During this part of the experiment participants were presented with 180 pairs of stimuli in three blocks of 60 stimulus pairs each. This allowed the participants to take a break between blocks in order to keep their attention focused on the experiment, and to prevent them from tiring. The length of each break could be

"Severe myopia contributed to Ron's inferiority complex."	*Different*
"Smash lightbulbs and their cash value will diminish to nothing."	*Same*
"Seamstresses attach zips with a thimble, needle, and thread."	*Different*
"Peter Piper picked a peck of pickled pepper."	*Same*

Figure 2.2: Practice sentences used in experiment

controlled by the participant, who chose when the experiment would resume by pressing the space bar.

The presentation of stimuli was randomised to control for any ordering effects, and each pair was presented in the same way. The first sentence in the pair was presented through the headphones while the computer displayed a white screen with the words "Sentence A" positioned in the center. Then a blank white screen accompanied by silence was presented for 200ms, after which the second sentence was presented through the headphones while the computer displayed the words "Sentence B". The blank screen in the middle of the two sentences served as a visual signal that the sentence was changing from A to B. This therefore signalled that the first sentence had ended, and that the participant should be prepared for the second half of the pair. After the pair had been presented the computer displayed a screen which asked "Were Sentence A and Sentence B the same or different in terms of naturalness?", and also presented a reminder that "0 = different, 1 = same". Following the participant's response, the experiment gave 800ms of silence before presenting the next pair. This was done to ensure that any sounds made by the keyboard as the participant entered his/her response did not interfere with the sound of the next sentence pair.

2.4 Analysis

2.4.1 Background

Many researchers have used perceptual experiments to obtain information about the quality of speech synthesis, and when analysing the results from such experiments, many have chosen to use multidimensional scaling. This technique has been widely used in areas where people have difficulty identifying the underlying parameters of perceived similarities or differences, such as in the study of tastes or smells (Grabe et al., 2003). As listeners often find it difficult to focus on just one aspect of synthesised speech, multidimensional scaling is an appropriate choice for speech quality analysis (King et al., 2005). As there is also poor understanding of how listeners perform auditory evaluations, this technique is often used to determine how many and which dimensions are being used in the listener's judgement (King et al., 2005). An introduction to multidimensional scaling is given below, followed by previous work involving multidimensional scaling as it is applied to perceptual experiments.

2.4.1.1 Multidimensional Scaling

Multidimensional scaling (MDS) is a mathematical procedure designed to show the "hidden structure" in a set of data (Barry et al., 2002). It is used to show how different or similar two objects or stimuli are perceived to be, as it outputs a geometric configuration where stimuli that are perceived as being similar are placed closer together, and those that are perceived as dissimilar are placed farther apart. The stimuli are capable of being arranged in one or more dimensions in order to find a configuration that best fits the data.

MDS places data points according to the predetermined dimensions into what is know as the "stimulus space" (Barry et al., 2002). It uses a function minimisation algorithm to move the objects around the stimulus space in order to pro-

duce a configuration that best approximates the perceived distances between the objects (Electronic Textbook - StatSoft, 2003). By moving and rearranging the objects in the stimulus space, the algorithm can maximise the goodness-of-fit, or how well a certain configuration reproduces the distances between the objects. The most common measure of the goodness-of-fit is the stress measure, although RSQ (r squared) can also be used (Kruskal and Wish, 1978).

The stress value essentially produces the sum of the squared deviations of the distances, which means that the smaller the stress value, the better the goodness-of-fit is (Kruskal and Wish, 1978). In general, the more dimensions used, the lower the stress value, and if the same number of dimensions as number of variables is used, the distances can be reproduced perfectly (Kruskal and Wish, 1978). However, using fewer dimensions allows the distances to be explained in a less complex manner.

The RSQ value represents the proportion of variation in the data that is accounted for by a particular scaling solution, meaning that the higher the RSQ value, the better the fit between the data and the scaling model (Barry et al., 2002). Therefore, a low stress value and a high RSQ value means a better goodness-of-fit.

When analysing the output from MDS, distances between the objects can be explained in terms of the dimensions, as it is theorised that the dimensions correspond to the attributes that participants use when judging the similarities and differences between stimuli (Hall, 2001). One dimension may signify intonation, for example, with natural intonation occurring at one end of the axis, and unnatural intonation at the other, while another dimension may represent the number of clicks or bad joins.

When it comes to speech analysis, paired comparisons procedures and MDS can be used to study the manner in which listeners weigh acoustic information (Allen and Scollie, 2002). Although only a few studies have used MDS to analyse auditory stimuli, it has been shown that MDS is able to determine what and how many acoustic features are perceptually important to a listener

(Barry et al., 2002; Allen and Scollie, 2002). Since it is difficult to determine which parameters participants pay attention to when judging the quality of synthetic speech, listeners can simply be asked to make judgements about the similarities or differences among speech samples without specifying which parameters they should base these judgements on (Lee et al., 2001; Stöber et al., 2001; Syrdal and Conkie, 2004). MDS analysis will then be able to determine the underlying perceptual features that the listeners used to make the judgements.

MDS has already been successfully applied to speech perception research such as the perception of differences in segmental structure and of voice quality effects (Barry et al., 2002; Iverson et al., 2003). It is also capable of providing insight into the subtle processes involved in judging synthetic speech, and so was an appropriate choice for this project (Barry et al., 2002).

2.4.1.2 Previous Work

As mentioned above, many researchers have used multidimensional scaling when analysing the results obtained from perceptual experiments. The following outlines previous work involving multidimensional scaling as it is applied to perceptual experiments.

Iverson et al. (2003) performed a perceptual experiment which tested participants' ability to distinguish between the English phonemes /r/ and /l/. The experiment was given to speakers of English, German, and Japanese in order to determine where their perceptual spaces were for each phoneme. Participants were presented with /ra/ and /la/ tokens in pairs and asked to rate the acoustic similarity of each pair. Multidimensional scaling was used to analyse the data from the experiment and map each participant's perceptual space by placing perceptually similar stimuli closer together, and dissimilar stimuli farther apart. The results from the MDS analysis then showed how the perceptual spaces for /r/ and /l/ were altered by native-language exposure. Japanese adults were most sensitive to the acoustic cue of F2, while English

and German speakers paid more attention to F3. This difference was made salient using MDS, which placed the stimuli in a two-dimensional Euclidean space which could be viewed easily.

Barry et al. (2002) conducted a perceptual experiment which tested tone discrimination in Cantonese. Groups of normal and hearing-impaired children were used in the experiment, which presented a background stimulus before changing it to a different stimulus. At the change of stimulus, participants were required to indicate they had perceived a difference. Multidimensional scaling was used to determine what differences, if any, existed between the information used by the different groups of children to distinguish tones. Through analysis of the MDS output it was shown that the most salient acoustic features in tone were average pitch height and pitch direction.

Allen and Scollie (2002) conducted a study to examine the effect of parameter value distributions on the similarity ratings obtained in a perceptual test. They first wanted to determine which acoustic characteristics were the most important to participants when they rated the similarity of stimuli. Next they wanted to determine whether or not this changed with variations to the distribution of values for an acoustic parameter. The two parameters which were manipulated were mean frequency and the number of components in the stimuli. Participants were asked to rate the similarity of pairs of tones and tone complexes. The similarity measures obtained were then analysed using multidimensional scaling. The results showed that frequency was the most important acoustic property.

Hall (2001) presents a study which aimed to develop a subjective speech quality test. First the advantages and disadvantages of commonly used quality rating methods were discussed, including Diagnostic Acceptability Measure (DAM), Mean Opinion Score (MOS), Degradation MOS (DMOS), and Comparison MOS (CMOS). A new speech quality test was then proposed that uses properties of multidimensional scaling. Before this was carried out, two perceptual experiments were done. The first required participants to rank the per-

ceptual distances between samples of speech processed by a variety of codecs. This was done by presenting participants with three stimuli and asking them to identify the two versions that sounded the most different. The results were then processed using MDS to obtain the stimulus space. In the second perceptual experiment, the perceptual attributes of the three dimensions of the stimulus space were identified. This established that different subjects use the same dimensions when judging the similarity of stimuli, but that they weight these dimensions differently. The three dimensions were identified as naturalness, noisiness, and the amount of low-frequency content.

Grabe et al. (2003) present a paper which investigates whether or not native language influences the perception of similarities and differences among intonation contours. A perceptual experiment was conducted with participants of different language backgrounds. An English utterance was synthesised many times, each with one of seven falling and four rising intonation contours. Native speakers of English, Iberian Spanish, and Mandarin Chinese all rated the differences between each pair of non-identical stimuli. The responses were then analysed using multidimensional scaling, which showed that the different language groups responded with statistically different perceptual configurations of the falling contours.

These works show that multidimensional scaling is able to provide valuable information regarding the parameters that listeners use when judging synthetic speech. It is an appropriate method for analysing results from perceptual experiments, such as listening experiments, as it identifies the underlying parameters of perceived similarities or differences.

2.4.2 Data Analysis

The responses collected in the two experiments were merged into a single file using E-Merge. Using E-DataAid and Excel, graphs were created showing each participant's response to each of the stimulus pairs, as well as the total

number of each response given for all of the pairs. Graphs were also made showing each participant's response to each of the four practice sentences. A series of difference matrices were then created showing the total number of "0" responses (indicating the participant perceived the two sentences as sounding different in terms of naturalness) for each of the pair combinations. This series was then condensed into one 9-by-9 matrix (the 9 parameter versions as shown in Chapter 2, Table 2.3) which produced a matrix containing cells for each possible parameter combination pair. Into these cells were then entered the total number of "0" responses for each of the stimulus pairs. This combined matrix was then entered into SPSS and analysed using multidimensional scaling. The multidimensional scaling was run using 1, 2, 3, 4, and 5 dimensions in order to find the number of dimensions which produced a configuration that best fit the data. The results of this analysis will be discussed in the following chapter.

2.5 Summary

This chapter discussed the experiment process which was carried out during the project. The weight values for the target costs were changed, and new sentences were synthesised to analyse the effect of the parameter changes. A total of 18 different parameter combinations were tested, and a ranking system was set up to determine which parameter combination had performed the best. A preliminary analysis was done based on this ranking, which determined the parameters that would be manipulated in the experiment. The experiment design was then discussed, detailing the materials, participants, and procedure used during the experiment. An introduction to multidimensional scaling was given, followed by a summary of previous work which showed that multidimensional scaling was an appropriate way of analysing speech perception. The results from the experiment were then analysed using MDS, and this analysis is discussed in the following chapter.

Chapter 3

Results and Conclusion

3.1 Introduction

This chapter includes the results obtained from the analysis of the data and discusses the possible meaning of these results. The data was analysed using multidimensional scaling, and a discussion of the output and each of the dimensions is given below. Finally, a conclusion is given, and ideas concerning future work and ways to improve the current experiment are discussed.

3.2 Results

The Excel graphs of the participants' responses were analysed in order to check that each participant had correctly answered the practice questions and could therefore be assumed to have understood the instructions for the experiment. Each of the participants answered half or more of the practice questions correctly, and so it was concluded that they had all understood the experiment enough to be included in the final analysis.

The data was copied into SPSS and analysed using multidimensional scaling. The results for the stress and RSQ values are shown in Table 3.1.

	Stress Value	RSQ
1 Dimension	0.227	0.895
2 Dimensions	0.144	0.936
3 Dimensions	0.094	0.962
4 Dimensions	0.076	0.971
5 Dimensions	0.059	0.980

Table 3.1: Stress and RSQ values from multidimensional scaling analysis

Barry et al. (2002) state that in order to properly interpret the output of MDS analysis, the true dimensionality must be determined. This can be done by plotting stress or RSQ against "Dimension" to determine the "elbow point" (Barry et al., 2002), which is the point in the graph where increasing the number of dimensions will only result in very slight improvements in the stress or RSQ values.

As shown in Table 3.1, using five dimensions resulted in the lowest stress value. However, I decided to do the analysis of the data using three dimensions. This is because the stress value does not change very much when the data is analysed using three, four, or five dimensions, but it is much easier to work in three dimensions rather than in four or five. Also, the biggest changes in the stress value occurred between the first and second dimensions, and between the second and third dimensions, showing that three dimensions provided a much better goodness-of-fit than just one or two dimensions, and that it did not increase significantly when four or five dimensions were used. This is illustrated in Figure 3.1 which shows that the so-called "elbow point" (Barry et al., 2002) occurs at three dimensions. Iverson et al. (2003) also state that the appropriate dimension should model at least 90% of the variance, and as can be seen in Table 3.1, using three dimensions allows more than 96% of the variance to be modelled. For these reasons the data was analysed using three dimensions.

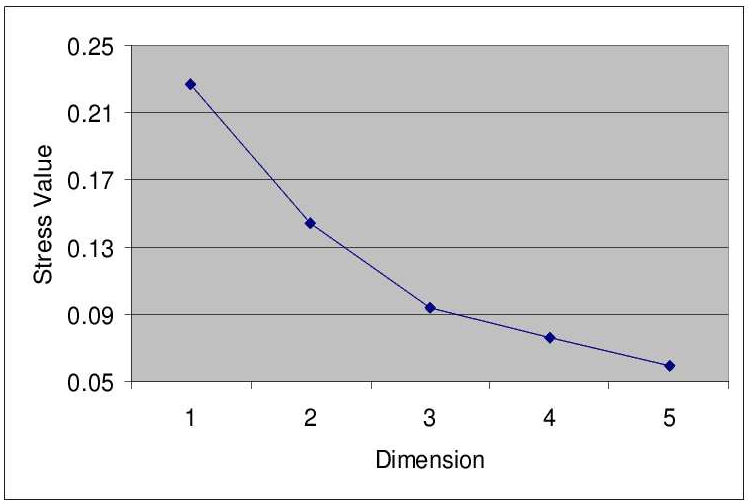

Figure 3.1: Stress values corresponding to each dimension

	Group 1			**Group 2**			**Group 3**		
	V9	*V1*	*V2*	*V3*	*V8*	*V5*	*V7*	*V4*	*V6*
Stress	1	1	1	5	10	10	10	5	10
Position in Syllable	10	5	10	1	1	1	10	10	5
Position in Phrase	10	10	5	10	10	5	1	1	1

Table 3.2: Groups of versions identified by MDS analysis

The results from SPSS showed that the data was divided into three distinct groups of three parameter versions each. A graph of the Euclidean distances between the versions is shown in Figure 3.2, which shows the groupings of the data points. For easier reference, the nine parameter versions used in the experiment are shown again in Table 3.2 in the groups identified by MDS analysis. Further discussion of the versions will correspond to these groups.

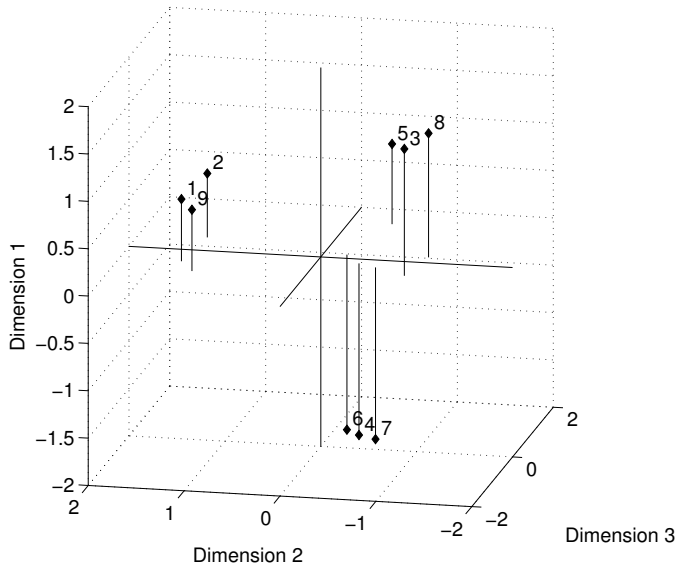

Figure 3.2: Euclidean distances between versions

3.3 Discussion

From Table 3.2 we can see that all versions in a group have given the weight value of 1 to the same parameter. There is no other common pattern for the versions in a group, and we can therefore conclude that groups are formed based on which parameter has been assigned the value of 1. This division into groups shows that participants perceived versions as being similar if they were synthesised with the weight value of 1 on the same parameter, and that versions which were perceived as less similar were synthesised with the weight value of 1 on different parameters. This suggests that assigning a small weight value to a parameter has a more perceptually noticeable effect on the resulting synthesis, and that listeners are more sensitive to a lack of weight on important parameters.

In order to determine if perceived similar versions were actually physically similar, a list of the diphones used in the synthesis of each sentence was created. This was done in Festival by re-synthesising the utterances and saving

each one separately. Then a list of relations was printed to the screen using the command "utt.relation.print utt 'Unit". This printed a list containing the name of each unit used, the target cost and join cost for the concatenation of each unit, and other information relating to the units used. A Python script was then written to extract the list of diphones used during synthesis. This list of diphones is shown at the end of the report in Appendix B.

The list of diphones showed that in general, the perceived groups corresponded to physical differences in the utterances. Versions that were grouped together did in fact tend to produce synthesised sentences using the same or similar diphones, while versions from other groups often used quite a different set of diphones for the synthesis of the same sentence. This shows that the parameter that had been given the value of 1 had a bigger influence over the selection of diphones chosen for synthesis than the parameters which had been given a value of 5 or 10. This suggests that a lack of weight on a particular parameter can have more effect on the selection of units than a large weight.

When one of the parameters in this experiment is given the weight of 1, it means that it is weighted the same as all the other parameters that were deemed less important in the initial analysis of parameters (discussed in Chapter 2). In other words, it means that the parameter is given a very small degree of importance. If this parameter is quite important in producing natural-sounding speech, the low weight assigned to it may result in poor synthesis. This is because all the other parameters are essentially given equal importance, and may therefore cause a selection of units which might normally have been regarded as quite unnatural. Without a large weight on an important parameter, it becomes less significant, and has less control over the sequence of units chosen. If this parameter is something which listeners pay particular attention to, a lack of consideration for this parameter may result in synthesis which is noticeably less natural-sounding. If, however, a parameter is given a large weight, for example 5, when all other parameters are given the small weight value of 1, it seems that it will not make much difference if the first parameter's weight is doubled to 10. This is because it is already marked as much more impor-

tant than the rest of the parameters, and unit selection will be strongly influenced by this single parameter. This means that an increase in weight on the parameter will make little difference, as the unit selection will still be based almost solely on that single parameter. Therefore, there might be a threshold of weight, which, once passed, begins to lose effect on the unit selection process, because once a parameter is weighted as the most important one, it may make no difference whether or not it is then given more weight, as it will remain the most important.

In order to determine what the perceived differences in the groups translated into, the three dimensions from the SPSS output were collapsed. This means that for each dimension the versions were plotted on a one-dimensional line for that dimension only, while ignoring the coordinate values for the other two dimensions. For each dimension this produced an ordered set of all the versions as they appeared on the graph, from the smallest coordinate value to the largest. This ordered list of collapsed dimensions is shown in Table 3.3.

Dimension 1	V6	V7	V4	V9	V1	V2	V5	V8	V3
Dimension 2	V8	V3	V7	V5	V4	V6	V9	V2	V1
Dimension 3	V9	V3	V1	V7	V4	V6	V8	V2	V5

Table 3.3: Collapsed dimension values

Using these collapsed dimensions, an auditory analysis was done in an attempt to determine what changed across the dimensions. In other words, for each dimension, I wanted to know why each version had been placed where it was in the graph. Were the versions at one end of a dimension clearer, more natural, or slower than versions at the other end? What was changing across each of the dimensions? By listening to versions of each sentence at one end of the dimension and comparing them to versions at the other end of the dimension, I attempted to uncover the reason each version had been placed where it was.

Unfortunately, this proved to be very difficult, as each of the sentences exhibited a different change, or no audible change at all across a dimension. The changes were subtle, and did not seem to happen smoothly across a dimension, but were instead abrupt. In other words, many of the changes were only audible at a certain point in the dimension. All versions on one side of the change sounded like one group, and the versions on the other side sounded like a different group, but within the groups the versions sounded the same. For this reason it was difficult to detect any sort of transition, when only an abrupt change was audible.

3.3.1 Dimension 1

The dimensions were analysed by trying to determine what each sentence was able to reveal about that dimension. A 2-dimensional graph of dimensions 1 and 2 is given in Figure 3.3, and shows the distribution of versions more easily than the 3-dimensional graph given previously. For easier reference, the sentences used in the experiment are also given again.

1. "Will you please describe the idiotic predicament."
2. "Only the most accomplished artists obtain popularity."
3. "Young people participate in athletic activities."
4. "Etiquette mandates compliance with existing regulations"
5. "Continental drift is a geological theory."

Table 3.4: Sentences used in the experiment

For the first dimension, I noticed a different phenomenon for each sentence. In sentence one, Groups 1 and 3 (refer to Table 3.2) sounded identical, while Group 2 had a definite hesitation on the /r/ of "predicament" which sounded quite unnatural. This was the only audible difference throughout the versions. Therefore, for sentence one it was unclear why there was a spread of versions

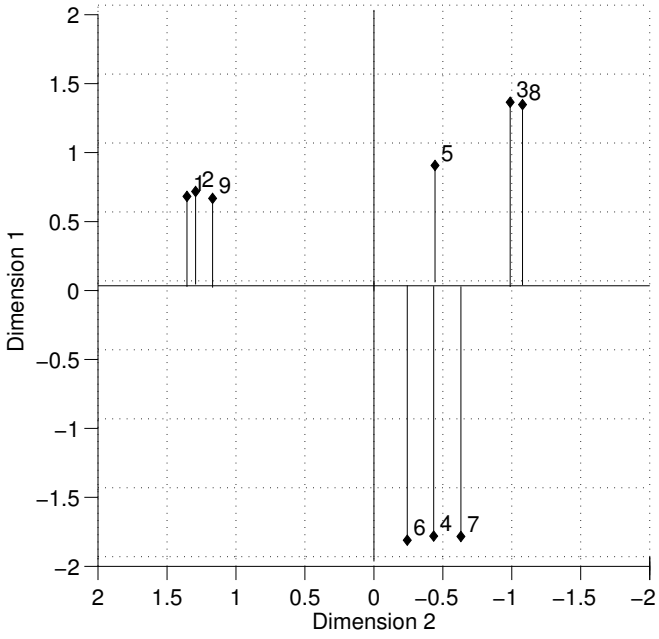

Figure 3.3: Distribution of versions in dimensions 1 and 2

across dimension one, as only two groups were audible. According to the perception of each sentence, Groups 1 and 3 should have been placed on the same point in the dimension, with the versions in Group 2 clustered together on a different point in the dimension. Although Group 2 sounds as though it should have been placed away from the others, when looking at Figure 3.3 it is clear that the versions in Group 3 are the outliers in dimension one. Therefore this sentence did not help to explain dimension one.

For sentence two I noticed only a small difference in the pronunciation of "obtain". Again Group 2 seemed to be separated from the other groups in this dimension, producing a pronunciation of "obtain" that sounded quite relaxed, with a reduced vowel for the "o", and no aspiration on the "t" (something like /əbteyn/). For Groups 1 and 3, the word "obtain" seemed more carefully pronounced, resulting in a pronunciation more like /obt^{h}eyn/. Again, there was perceptually no difference within Groups 1 and 3, and within Group 2,

and when compared with the list of diphones used it was shown that there was in fact almost no difference in the sequence of diphones chosen in Groups 1 and 3, and within Group 2. Again it was unclear why dimension one had been plotted as it had, as up until now it seemed that Group 2 contained the versions which should be most separated from the others.

Sentence three proved inconclusive, as no difference was detected between the first version and the last. This time however it was Group 1 which differentiated itself as a group, although it was located in the middle of the dimension. A subtle perceptual difference occurred in these three versions which sounded as though pronunciation were slightly clearer. This was only noticeable on the word "athletics", and manifested itself as a slightly more aspirated /t/. This was supported by the list of diphones, which showed that the two groups at either end of the dimension were virtually identical, while all the versions in Group 1 were quite different from those in Groups 2 and 3, but were exactly the same as each other.

Sentence four was perceived as being segmented into three definite groups, divided by the first word of the sentence, "Etiquette". When moving across the dimension from smallest value to largest, the pronunciation of "Etiquette" improved. The division into the three groups as shown in Tables 3.3 and 3.2 was quite easily detectable, with Group 2 recognisable as containing the worst renditions of "Etiquette". More than the others, this sentence showed that there was a perceptual segmentation of the versions into the three groups. Previous sentences had shown only two groups, with three versions being set aside as the outliers. While the previous three sentences did alternate which three versions were different from the other six, this sentence was the first to show that the versions could be split into the three groups in a single sentence. When this perceptual grouping was compared to the list of diphones it was shown that the three groups were in fact very different from each other, and that members of a group were very similar to each other. Therefore it was not only easy to perceptually divide the versions into three groups, but also physically, in terms of the diphones chosen for synthesis.

In sentence five there was a very distinct and very audible difference which separated the versions in Group 3 as the outliers. All other versions were perceptually indistinguishable, and when the diphone list was consulted it was discovered that they were also almost completely physically identical. Group 3, however, differed from the rest on the last four diphones in the sentence. This difference in only four diphones turned out to be quite critical for the separation of the group. Perceptually, these four diphones created a strange rising intonation at the end of the sentence, which sounded very unnatural, as normal human speech tends to decline in intonation at the end of a sentence (Ladd, 1996). Versions with this decline in intonation were strongly perceived as being different from the other versions, as is shown in the difference matrix entered into SPSS. This matrix showed that 90% of the time participants heard Group 3 as being different from the other versions (325 responses out of 360 times that versions 4, 6, and 7 were paired with one of the other six versions). This shows that Group 3 was very distinct, and that it was strongly and almost unanimously perceived. It produced versions that are so strongly perceived as different, that almost every time they were presented with another version they were indicated as sounding different. This strong agreement is why Group 3 appears as such an outlier on dimension one.

Interestingly, the parameter which was set at 1 for Group 3 was Position in Phrase, meaning that this parameter got very little weight. As mentioned before, intonation tends to decline at the end of a sentence or phrase, and Position in Phrase would be responsible for choosing units from the database which were in the correct place in the phrase, or in other words, with the correct intonation declination. When this parameter was given a small weight, it produced synthesis that had an obvious unnatural quality, which is reflected in the strong sense of difference perceived by the participants. This suggests that Position in Phrase is an important parameter in producing natural-sounding synthesis, as failure to place importance on this parameter can result in very unnatural-sounding intonation. The strong agreement between participants indicating the difference of these three versions also suggests that intonation is

a component of synthesis which can have a large and important effect on the naturalness of the synthesis. It suggests that participants are quite sensitive to changes in intonation, and that failure to comply with the standard of declination at the end of a sentence is perceived as a great detriment to the quality of the synthesis.

3.3.2 Dimension 2

As in the first dimension, it was unclear which one phenomenon was changing, as each sentence seemed to show something different. For the first sentence it was again the slight hesitation on the /r/ of "predicament" that was perceptually different, and the pronunciation got better when moving along the dimension from smallest value to largest. Both perceptually and physically it is Group 2 which is different, containing this unnatural-sounding hesitation, but in the graph it is Group 1 which is more separate from the rest. In fact, Group 2 does not stand out at all in the dimension. Therefore sentence one does not explain why Group 1 appears as an outlier on this dimension, as Group 2 seems to be the group which is most different.

Sentence two again shows a difference in the pronunciation of "obtain". This time the pronunciation becomes clearer and more pronounced as the values for the coordinates increases, but again it is Group 2 which appears different from the other two groups. When the list of diphones is consulted it shows that Groups 1 and 3 are almost identical in terms of the sequence of units chosen, therefore sentence two again does not explain why Group 1 is separated from the other versions in this dimension.

Sentence three showed something which was not quite noticeable in the first dimension. This sentence revealed that Group 1 contained a clearer articulation of the word "athletics". For this sentence it is clearly Group 1 which is different from the other two, and this is supported by the diphone list which shows that the other two groups are different than Group 1, but virtually iden-

tical to each other. In this dimension the articulation of "athletics" becomes clearer as the coordinate values increase. These three versions were in the middle of the first dimension (as shown in Table 3.3) and so dimension one did not show this difference well. It is only in the second dimension that this difference becomes clear. This appears to be why Group 1 is separated from the rest of the versions in the dimension. Perhaps this dimension is plotted especially to make this difference clear and to show that Group 1 is best explained as its own group, and should not simply be grouped with the other versions.

Dimension two did not show the differences in sentence four as clearly as dimension one had. Sentence four showed the three groups of versions based on the pronunciation of "Etiquette", and this is shown clearly in dimension one which places versions on the dimension in order of best pronunciation of "Etiquette" to worst. Dimension two however plots Group 3 in the middle of the dimension, when they are clearly the best examples of the word "Etiquette". Therefore, this dimension does put the worst versions of the word at one end of the dimension, and better versions at the other, but the best examples of "Etiquette" are in the middle. This does not show a gradual increase or decrease in naturalness, but instead just three distinct groupings, whose place in the dimension does not seem obvious. Therefore it is unclear what is happening in this dimension for sentence four. Perhaps the dimension was created to show the separation of Group 1, which was identified in the previous sentence.

Sentence five was similar to sentence four in that it did not give a clear indication of what was going on across the dimension. This sentence contained the most audible difference between versions out of any of the sentences - a strange rise in intonation at the end of Group 3. This group was placed in the middle of the dimension, however, and the reason for this is not apparent. As with sentence four, this dimension did not show the differences in this sentence as well as dimension one, as the three versions at one end of the dimension sounded the same as the three versions at the other end. It was only in the middle of the dimension that I could notice a difference perceptually.

3.3.3 Dimension 3

A 2-dimensional graph showing dimensions 1 and 3 is given in Figure 3.4. This graph shows the distribution of versions across dimensions 1 and 3 more easily than the 3-dimensional graph shown previously.

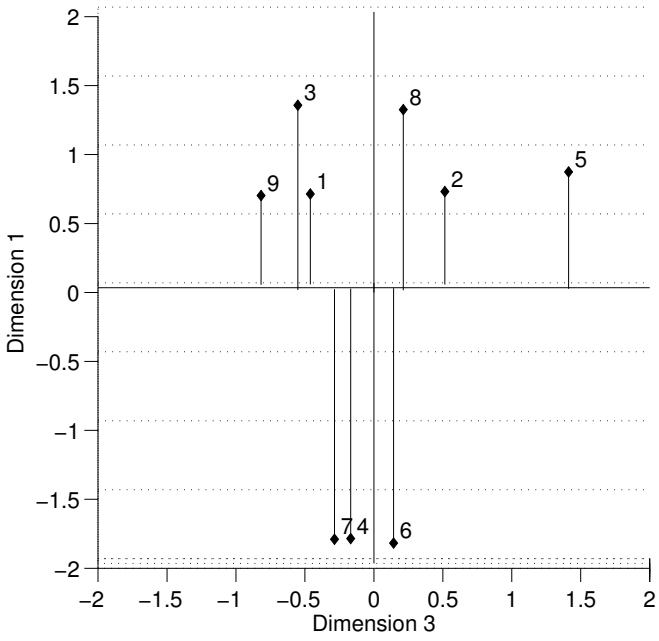

Figure 3.4: Distribution of versions in dimensions 1 and 3

From the graph in Figure 3.4 we can see that there does not seem to be any sort of group set apart from the rest. Version 5 is placed slightly farther away from the rest of the versions, although the previous groupings shown in the graphs above are not apparent. Although the versions in Group 3 are still close together, the other versions previously grouped together are quite spread out and overlap with each other. Once again the five sentences were investigated in an attempt to find any information that could be used to explain dimension three.

Sentence one did not give any hints as to what was changing across dimension three. As before, the only audible difference was the hesitation of the /r/

in "predicament", which occurred in all versions in Group 2. Unfortunately, these three versions were quite spread out across the dimension, and therefore did not allow any perceivable difference to become apparent across the dimension. If versions 2 and 3 had switched places, it would have seemed as if the pronunciation of "predicament" was getting worse. However, these versions are quite far apart and five other versions occur between them. Therefore, with the versions in Group 2 scattered seemingly arbitrarily throughout the dimension, the conclusion for this sentence is that is does not show anything about what is happening in the 3rd dimension. Also, in this sentence versions 3 and 5 are identical in terms of their diphone sequence, and so it does not explain why version 5 is separated from the rest of the versions.

The analysis from sentence two was quite similar to that of sentence one. Again it was Group 2 that was perceptually different from the other six versions, but again there did not seem to be any pattern. Because the other six versions were perceptually indistinguishable, it seemed as if the versions in Group 2 should have been grouped together again. Versions 3 and 5 were again identical, so the distance between version 5 and the rest of the versions was quite strange. The distribution throughout the dimension again made it too difficult to conclude what was happening across the dimension.

In sentence three it was Group 1 that was different, but again this was not plotted in the dimension in any sort of group. Version 5 was identical to versions 3, 4, 6, and 8, but this is not reflected in the graph. Sentence four was equally unhelpful, as all versions of "Etiquette" seemed randomly placed in the dimension with no clear groupings at all. Sentence 5 was the same as in dimension two, having the versions with the unnatural intonation in the middle of the dimension and leaving the two ends of the dimension sounding identical. Therefore after listening to all five sentences it was still unclear what was happening in dimension three.

3.3.4 The Versions as Groups

After analysing what was happening across dimensions, I now wanted to know what was happening within the groups. The list of diphones used in the synthesis of each utterance was consulted to determine what effects each set of parameters had on the sequence of diphones chosen.

As mentioned previously, the groups seem to have been chosen based on which parameter had been assigned the value of 1. In other words, all members of a group had the weight value of 1 on the same parameter. This low weight had the biggest impact when it came to choosing diphones for synthesis, as versions in different groups often had quite different sequences of diphones, but versions within a group were usually very similar or identical.

For sentence one, the parameter settings for Group 2 made the most difference to the diphones chosen for synthesis, while the other two groups were quite similar. Group 2 had the value of 1 on the Position in Syllable parameter, which therefore suggests that when Stress or Position in Phrase are set to 1, there is little effect on the end result of the synthesised sentence. However, when Position in Syllable is given such little importance, the effect on the final synthesis of the sentence is much greater, and much more audible. Group 2 was also the least natural-sounding, which suggests that Position in Syllable is an important parameter, and that when it is given a small amount of importance, the resulting synthesis sounds unnatural. As these three versions produced a strange pronunciation of the /r/ in "predicament", it suggests that Position in Syllable is important for choosing units which are from the correct place in the syllable, and therefore the correct length. Phonemes which are at different places in the syllable may be different lengths, and this may be the reason that such hesitation on the /r/ occurred in these three versions.

Sentence two shows a similar grouping phenomenon, where again the parameter settings for Group 2 make the most difference in terms of the diphones chosen. As before, it does not seem to make a difference what value Stress

or Position in Phrase is given, as the resulting synthesis from their varied values is almost identical. For this sentence however, the difference in Group 2 is much smaller, and does not make such an audible difference. Also, although a difference is detectable, it is difficult to say if it causes the versions to sound more or less natural. Interestingly, the difference occurring in Group 2 again seems to be in the length of the phonemes, as it produces a pronunciation of "obtain" which has a shorter-sounding /o/. Fortunately this difference does not lead to unnatural-sounding synthesis, and instead perhaps gives the impression of less careful speech. Therefore for this sentence it is unclear which groups sound the most or least natural. It is clear, however, that Position in Syllable has once again had the most effect on the resulting synthesis, but only when it was given a low weight value.

Sentence three showed something very different from the previous two sentences. In this sentence it is the parameter combinations in Group 1 that made the most difference to the sequence of units chosen. These three versions assigned the value of 1 to the Stress parameter, and produced sequences of units that were identical to each other. In this sentence, however, the parameter combinations for the versions in Groups 2 and 3 produced identical sequences of units, which implies that Position in Syllable is not always the most important parameter. Again, though a small difference between the groups was audible, it was difficult to say which one was more natural-sounding. Although one could be said to sound more articulated, this did not necessarily imply that it sounded more natural. In this sentence Position in Syllable seemed to have no effect, and the group which assigned this parameter the value of 1 produced the same sequence of units as the group which assigned Position in Phrase the value of 1. This sentence shows that Stress can also have a large effect on the sequence of units chosen, and that the answer to which variation of parameter values is the most important varies according to which sentence is used to test the parameters. While the first two sentences had begun to suggest that Position in Syllable was the most sensitive to changes in weight value, this sentence shows that different parameters may be more important for different

sentences. This also means that deciding which parameter variation is most important may be unachievable, as different parameters are important for different sequences of units.

Sentence four showed that each of the groups had chosen very different sequences of diphones. This further complicates the problem of finding the most important variation of parameter weights by showing that each group was affected quite strongly by the changes made to the parameter values. This means that all three of the parameters have the potential to largely alter the sequence of diphones chosen when they are given a small weight. Group 1 had assigned Stress the value of 1, and produced a believable pronunciation of "Etiquette" and a natural-sounding intonation. The natural intonation may have been a result of the high weight on Position in Phrase, which is responsible for choosing diphones from the correct place in the phrase (and therefore diphones with better declination). Group 3 produced the clearest version of "Etiquette", but had a slightly strange intonation. This group had given Position in Phrase the value of 1, which suggests that the slightly strange intonation is a result of the lack of weight given to Position in Phrase, causing diphones from the wrong place in the phrase to be chosen. Group 2 produced a very bad and almost incomprehensible pronunciation of "Etiquette", and had given Position in Syllable the value of 1. This lack of weight may have resulted in diphones from the wrong place in the syllable being chosen, which may have been the wrong length and could have caused the unnatural-sounding synthesis. It is interesting, however, that the large weight on Stress in versions 5 and 8 did not manage to overcome the lack of weight on Position in Syllable, as "Etiquette" is stressed on the initial syllable, and this is where the problem occurred. This means that because of the high weight on Stress, the sequence of diphones should have included a phoneme for the first sound in "Etiquette" which had come from a stressed word, and should have sounded more natural. This sentence therefore suggests that a low weight on Position in Syllable may have more effect on the pronunciation of a word than a high weight on Stress.

In sentence five it was the parameter combinations for Group 3 that had the most effect on the sequence of units chosen for synthesis. This group had given the value of 1 to Position in Phrase, which produced versions of the sentence that only differed from the other two groups in the last four diphones. In fact, all nine versions were physically very similar, which shows that it is possible for none of the parameter combinations to have very much effect on the resulting synthesis. Even though the parameter combinations for Group 3 did not have much effect on the diphones chosen, the small physical effect that it did have turned out to be a very large effect perceptually. As mentioned before, assigning the value of 1 to Position in Phrase resulted in a very audible abnormality in the intonation of Group 3. This is because without enough weight on Position in Phrase, diphones which were originally in the middle of a sentence when recorded (higher in pitch than at the end) may be placed at the end of the sentence during synthesis, resulting in a strange pitch jump.

A list of the target costs and join costs for each sentence was also created and analysed in terms of the groups found by the MDS output. Unfortunately no pattern was found, as no version consistently produced the lowest, or even the highest cost. At times a low cost did correspond to an utterance that sounded natural, but this was not always the case. In fact, for sentence five, the two versions with the lowest join cost were two which contained the strange intonation rise.

3.3.5 Summary of Results

As it was quite difficult to determine what was happening across the three dimensions, it appears as if the dimensions only reflected why the versions were segmented into three groups. Dimension one at first appeared confusing, as in the beginning of the analysis it appeared to be Group 2 which was different from the other versions. Sentence five, however, showed that dimension one identified Group 3 as different from the other versions. Dimension one clearly shows these three versions quite far away from all the others, and sentence

five showed why this had been done. It presented a very audible difference between these three versions and the rest, and explains why dimension one is plotted as it is.

The analysis of dimension two again started out unclear but became clearer with sentence three, which showed that Group 1 was a separate group. This dimension shows Group 1 as being quite separate from the rest, and also serves to create the versions in Group 2 as their own group. Although this is not clear simply from the dimension, which plots Group 2 quite close to Group 3, dimension one has already clustered Group 3 into its own group. Therefore the separation of Group 1 does not only identify it as its own group, but also essentially groups the versions in Group 2 together, as they now belong to neither Group 3 nor Group 1. Therefore dimension two also serves to cluster Group 2 together by showing that they are not part of Group 1. Dimension three on the other hand did not appear to show any new groupings or changes. This suggests that it may show differences which are not immediately obvious, and which are quite subtle.

The difficulty in determining what was happening in the three dimensions shows that it was better to use three dimensions rather than four or five. Because there was such difficulty determining what was happening in only three dimensions, adding another, or even two more dimensions would have only made things more difficult. The subtle differences that caused the versions to be grouped as they were showed that multidimensional scaling is able to detect groupings that may not be immediately obvious. MDS is therefore an appropriate way of analysing perceptual data concerning the quality of synthetic speech, as it is able to model differences between utterances with only small audible differences.

The analysis showed that Group 2 produced the worst synthesis, as it was the worst group in three out of the five sentences. This group had given Position in Syllable a very low weight, and shows that the loss of this parameter often results in poor quality synthesis. Group 3 contained one sentence with a

very audible intonation discontinuity, and had given Position in Phrase a low weight. This suggests that Position in Phrase is important for controlling intonation contours throughout the utterance. Therefore, Group 1 was chosen as the group which produced the most consistently natural-sounding speech. This group had given Stress a low weight value, and showed that even without a large amount of importance on Stress, the resulting synthesis was still judged as the best quality synthesis in the experiment. This suggests that Stress is not as important to the quality of speech synthesis as Position in Phrase or Position in Syllable.

3.4 Future Work and Room for Improvement

Given more time, I would like to continue this experiment with a greater variety of sentences and parameter combinations. Because of the time constraints for this project, only 20 participants were able to take part in the experiment. Because each participant can only listen to a limited number of paired comparisons, in order to allow all versions of pairs to be heard I had to limit the stimuli to be made up of five sentences. Also, because of the limited number of stimuli I was able to use for the experiment, only nine different parameter combinations were tested. Ideally I would be able to test more combinations and sentences, as this would provide even more information regarding the acoustic parameters used in the judgement of speech quality.

I would also like to run a similar experiment designed to determine what each dimension represents. Hall (2001) conducted such an experiment, which presented two versions from the ends of the dimensions and asked participants what they thought was changing across the dimension. This is the way that I attempted to identify the dimensions myself, but an experiment would be a better way to provide the true identification of each dimension.

It would also be useful to run a follow-up experiment to show that the new parameter weights do produce better quality speech than the original parame-

ter weights. This could be done in the same manner as the current experiment, but with all the weight values set to their proper values, instead of having the value of 1. Synthetic speech using these parameter values could then be compared to synthetic speech using the original values.

Also, as mentioned previously, I developed a small hypothesis that a lack of weight on Position in Syllable sometimes caused the system to choose diphones that were too short. It would be interesting to listen to the actual diphone used in the synthesis and compare it to other diphones to determine if my hypothesis was correct.

3.5 Conclusion

The evaluation of synthetic speech has proved to be a difficult and interesting task. By conducting a perceptual experiment involving paired comparisons, this project aimed to contribute to current speech quality research by uncovering information about the relationship between target costs and the perceived discontinuity of synthetic speech. The goal of the project was to determine a better set of target cost weight values for the Festival speech synthesis program, allowing it to produce more natural-sounding synthesis.

The results from the experiment showed that when judging synthetic speech, participants pay attention to Position in Phrase, Position in Syllable, and Stress parameters. It was also found that participants pay close attention to intonation, but this is not a component of the target costs and is taken care of by join costs. The importance of these acoustic parameters was shown by the fact that participants grouped the stimuli based on which of these parameters was given the weight value of 1, which also reveals that listeners are more sensitive to the lack of weight on these parameters than to a large weight. The experiment results also showed that a lack of weight on these parameters also has more effect on the selection of units from the database than a large amount of

weight, as different versions which assigned the value of 1 to the same parameter produced similar diphone sequences.

Through analysis of the results it was shown that Position in Syllable was the most important parameter for high quality speech, as a lack of weight on this parameter produced what was perceived as the worst quality synthesis. Position in Phrase was also found to be quite important, as a lack of weight on this parameter produced a very noticeably bad sentence. Stress was shown to be less important, as a lack of weight on this parameter did not produce sentences that were perceived to be of particularly bad quality. Therefore, a better set of target cost weight values for the Festival speech synthesis program would give a large amount of weight to Position in Syllable, Position in Phrase, and Stress, as these three parameters are associated with good quality speech synthesis.

Appendix A

List of Sentences

1. “Peter Piper picked a peck of pickled peppers.”
2. “She sells sea shells by the sea shore.”
3. “Coconut cream pie makes a lovely dessert.”
4. “Where were you while we were away?”
5. “Critical equipment needs proper maintenance.”
6. “Artificial intelligence is for real.”
7. “Robin will allow a rare lie.”
8. “While waiting for Chipper she criss-crossed the square many times.”
9. “The altruistic dowager helped many malnourished vagrants.”
10. “Severe myopia contributed to Ron’s inferiority complex.”
11. “Seamstresses attach zips with a thimble, needle, and thread.”
12. “Smash lightbulbs and their cash value will diminish to nothing.”
13. “The 8:45 train to Liverpool Lime Street will depart from platform seven in fifteen minutes.”
14. “Comprehension usually precedes production. Quite often contextual cues are strong enough for the child to get the gist of an utterance without perhaps being able to understand the details.”

15. "There's a KLM flight arriving Brussels at ten to five. But business class is not available, and you'd need to connect in Amsterdam. If you want to fly direct, there's a BMI flight that arrives at 4:10, but it has no availability in business class either. There are seats in business class on the British Airways flight that arrives at 4:20. It requires a connection in Manchester though."
16. "This is a type of broach that was popular around the 1960's. It might not be instantly recognisable as jewellery, but it is important to remember that jewellery doesn't have to be expensive or elaborately crafted. Indeed the term jewellery encompasses an extraordinary range of accessories which people have used to decorate themselves."
17. "In Sioux Falls today, partly cloudy skies with periods of sunshine. Tomorrow, and Friday, mostly cloudy and cloudy skies. Saturday, partly cloudy skies with periods of sunshine. Sunday mostly sunny skies."

Appendix B

List of Diphones

The list of diphones used for synthesis is shown below. The nine parameter combinations are arranged in the groupings determined by multidimensional scaling. Diphones

Sentence 1 - "Will you please describe the idiotic predicament."

	1	2	9	3	5	8	4	6	7
1	nina_c2_018	nina_c2_018	nina_c2_018	nina_c2_018	nina_c2_018	nina_c2_018	nina_c2_018	nina_c2_018	nina_c2_018
2	nina_01_108	nina_01_108	nina_01_108	nina_01_108	nina_01_108	nina_01_108	nina_01_108	nina_01_108	nina_01_108
3	nina_01_108	nina_01_108	nina_01_108	nina_01_108	nina_01_108	nina_01_108	nina_01_108	nina_01_108	nina_01_108
4	nina_01_108	nina_01_108	nina_01_108	nina_01_108	nina_01_108	nina_01_108	nina_01_108	nina_01_108	nina_01_108
5	nina_01_108	nina_01_108	nina_01_108	nina_01_108	nina_01_108	nina_01_108	nina_01_108	nina_01_108	nina_01_108
6	nina_01_001	nina_01_001	nina_01_001	nina_01_001	nina_01_001	nina_01_001	nina_01_001	nina_01_001	nina_01_001
7	nina_24_075	nina_24_075	nina_24_075	nina_24_075	nina_24_075	nina_24_075	nina_24_075	nina_24_075	nina_24_075
8	nina_24_075	nina_24_075	nina_24_075	nina_24_075	nina_24_075	nina_24_075	nina_24_075	nina_24_075	nina_24_075
9	nina_04_014	nina_04_014	nina_04_014	nina_04_014	nina_04_014	nina_04_014	nina_04_014	nina_04_014	nina_04_014
10	nina_03_022	nina_03_022	nina_03_022	nina_03_022	nina_03_022	nina_03_022	nina_03_022	nina_03_022	nina_03_022
11	nina_03_022	nina_03_022	nina_03_022	nina_03_022	nina_03_022	nina_03_022	nina_03_022	nina_03_022	nina_03_022
12	nina_03_022	nina_03_022	nina_03_022	nina_03_022	nina_03_022	nina_03_022	nina_03_022	nina_03_022	nina_03_022
13	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049
14	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049
15	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049
16	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049
17	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049
18	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049	nina_r1_049
19	nina_16_075	nina_16_075	nina_16_075	nina_16_075	nina_16_075	nina_16_075	nina_16_075	nina_16_075	nina_16_075
20	nina_17_045	nina_17_045	nina_17_045	nina_17_045	nina_17_045	nina_17_045	nina_17_045	nina_17_045	nina_17_045
21	nina_17_045	nina_17_045	nina_17_045	nina_17_045	nina_17_045	nina_17_045	nina_17_045	nina_17_045	nina_17_045
22	nina_05_056	nina_05_056	nina_05_056	nina_05_056	nina_05_056	nina_05_056	nina_05_056	nina_05_056	nina_05_056
23	nina_07_074	nina_18_006	nina_18_006	nina_07_074	nina_07_074	nina_07_074	nina_18_006	nina_18_006	nina_18_006
24	nina_25_052	nina_18_006	nina_18_006	nina_07_074	nina_07_074	nina_07_074	nina_18_006	nina_18_006	nina_18_006
25	nina_25_001	nina_07_008	nina_07_008	nina_04_014	nina_04_014	nina_04_014	nina_07_008	nina_07_008	nina_07_008
26	nina_25_001	nina_04_014	nina_04_014	nina_04_014	nina_04_014	nina_04_014	nina_04_014	nina_04_014	nina_04_014
27	nina_02_068	nina_04_014	nina_04_014	nina_04_014	nina_04_014	nina_04_014	nina_04_014	nina_04_014	nina_04_014
28	nina_02_068	nina_17_042	nina_17_042	nina_19_001	nina_19_001	nina_04_014	nina_17_042	nina_17_042	nina_17_042
29	nina_23_083	nina_17_042	nina_17_042	nina_25_045	nina_25_045	nina_04_087	nina_17_042	nina_17_042	nina_17_042
30	nina_07_095	nina_21_016	nina_21_016	nina_14_049	nina_14_049	nina_14_053	nina_14_053	nina_14_049	nina_14_053
31	nina_07_095	nina_r1_015	nina_r1_015	nina_14_049	nina_14_049	nina_14_053	nina_02_071	nina_03_035	nina_13_048
32	nina_12_002	nina_12_022	nina_12_022	nina_07_027	nina_07_027	nina_11_035	nina_13_025	nina_10_028	nina_19_017
33	nina_25_096	nina_16_010	nina_16_010	nina_24_099	nina_24_099	nina_12_011	nina_16_010	nina_24_099	nina_07_100
34	nina_25_096	nina_07_067	nina_07_067	nina_24_099	nina_24_099	nina_12_011	nina_07_067	nina_24_099	nina_07_067
35	nina_25_096	nina_07_067	nina_07_067	nina_24_099	nina_24_099	nina_12_011	nina_07_067	nina_24_099	nina_07_067
36	nina_25_096	nina_07_059	nina_07_059	nina_24_099	nina_24_099	nina_12_011	nina_07_059	nina_24_099	nina_07_059
37	nina_25_096	nina_07_059	nina_07_059	nina_24_099	nina_24_099	nina_12_011	nina_07_059	nina_24_099	nina_07_059

Sentence 2 - "Only the most accomplished artists obtain popularity."

	1	2	9	3	5	8	4	6	7
1	nina_01_084	nina_01_084	nina_01_084	nina_01_084	nina_01_084	nina_01_084	nina_01_084	nina_01_084	nina_01_084
2	nina_01_084	nina_01_084	nina_01_084	nina_01_084	nina_01_084	nina_01_084	nina_01_084	nina_01_084	nina_01_084
3	nina_01_084	nina_01_084	nina_01_084	nina_01_084	nina_01_084	nina_01_084	nina_01_084	nina_01_084	nina_01_084
4	nina_01_084	nina_01_084	nina_01_084	nina_01_084	nina_01_084	nina_01_084	nina_01_084	nina_01_084	nina_01_084
5	nina_c2_014	nina_c2_014	nina_c2_014	nina_c2_014	nina_c2_014	nina_c2_014	nina_c2_014	nina_c2_014	nina_c2_014
6	nina_19_022	nina_19_022	nina_19_022	nina_19_022	nina_19_022	nina_19_022	nina_19_022	nina_19_022	nina_19_022
7	nina_19_022	nina_19_022	nina_19_022	nina_19_022	nina_19_022	nina_19_022	nina_19_022	nina_19_022	nina_19_022
8	nina_13_031	nina_13_031	nina_13_031	nina_13_031	nina_13_031	nina_13_031	nina_13_031	nina_13_031	nina_13_031
9	nina_13_031	nina_13_031	nina_13_031	nina_13_031	nina_13_031	nina_13_031	nina_13_031	nina_13_031	nina_13_031
10	nina_13_031	nina_13_031	nina_13_031	nina_13_031	nina_13_031	nina_13_031	nina_13_031	nina_13_031	nina_13_031
11	nina_13_031	nina_13_031	nina_13_031	nina_13_031	nina_13_031	nina_13_031	nina_13_031	nina_13_031	nina_13_031
12	nina_25_025	nina_25_025	nina_05_058	nina_25_025	nina_25_025	nina_05_058	nina_25_025	nina_25_025	nina_05_058
13	nina_17_010	nina_17_010	nina_23_046	nina_17_010	nina_17_010	nina_23_046	nina_17_010	nina_17_010	nina_23_046
14	nina_17_010	nina_17_010	nina_23_046	nina_17_010	nina_17_010	nina_23_046	nina_17_010	nina_17_010	nina_23_046
15	nina_17_010	nina_17_010	nina_23_046	nina_17_010	nina_17_010	nina_22_044	nina_17_010	nina_17_010	nina_23_046
16	nina_12_060	nina_12_060	nina_07_015	nina_12_060	nina_12_060	nina_22_044	nina_12_060	nina_12_060	nina_17_100
17	nina_22_047	nina_22_047	nina_22_047	nina_22_047	nina_22_047	nina_16_038	nina_22_047	nina_22_047	nina_22_047
18	nina_22_047	nina_22_047	nina_22_047	nina_22_047	nina_22_047	nina_16_038	nina_22_047	nina_22_047	nina_22_047
19	nina_22_047	nina_22_047	nina_22_047	nina_22_047	nina_22_047	nina_22_047	nina_22_047	nina_22_047	nina_22_047
20	nina_22_047	nina_22_047	nina_22_047	nina_22_047	nina_22_047	nina_22_047	nina_22_047	nina_22_047	nina_22_047
21	nina_22_047	nina_22_047	nina_22_047	nina_22_047	nina_22_047	nina_22_047	nina_22_047	nina_22_047	nina_22_047
22	nina_23_045	nina_23_045	nina_23_045	nina_23_045	nina_23_045	nina_23_045	nina_23_045	nina_23_045	nina_23_045
23	nina_23_045	nina_23_045	nina_23_045	nina_23_045	nina_23_045	nina_23_045	nina_23_045	nina_23_045	nina_23_045
24	nina_23_045	nina_23_045	nina_23_045	nina_23_045	nina_23_045	nina_23_045	nina_23_045	nina_23_045	nina_23_045
25	nina_23_045	nina_23_045	nina_23_045	nina_23_045	nina_23_045	nina_23_045	nina_23_045	nina_23_045	nina_23_045
26	nina_17_099	nina_17_099	nina_17_099	nina_09_050	nina_09_050	nina_09_050	nina_17_099	nina_17_099	nina_17_099
27	nina_09_053	nina_09_053	nina_09_053	nina_09_050	nina_09_050	nina_09_050	nina_09_053	nina_09_053	nina_09_053
28	nina_09_053	nina_09_053	nina_09_053	nina_17_083	nina_17_083	nina_17_083	nina_09_053	nina_09_053	nina_09_053
29	nina_09_053	nina_09_053	nina_09_053	nina_17_083	nina_17_083	nina_17_083	nina_09_053	nina_09_053	nina_09_053
30	nina_09_053	nina_09_053	nina_09_053	nina_17_083	nina_17_083	nina_17_083	nina_09_053	nina_09_053	nina_09_053
31	nina_22_004	nina_22_004	nina_22_004	nina_22_004	nina_22_004	nina_22_004	nina_22_004	nina_22_004	nina_22_004
32	nina_22_016	nina_22_016	nina_22_016	nina_22_016	nina_22_016	nina_22_016	nina_22_016	nina_22_016	nina_22_016
33	nina_22_016	nina_22_016	nina_22_016	nina_22_016	nina_22_016	nina_22_016	nina_22_016	nina_22_016	nina_22_016
34	nina_22_016	nina_22_016	nina_22_016	nina_22_016	nina_22_016	nina_22_016	nina_22_016	nina_22_016	nina_22_016
35	nina_22_016	nina_22_016	nina_22_016	nina_22_016	nina_22_016	nina_22_016	nina_22_016	nina_22_016	nina_22_016
36	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025
37	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025
38	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025
39	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025
40	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025	nina_c1_025
41	nina_17_057	nina_17_057	nina_17_057	nina_17_057	nina_17_057	nina_17_057	nina_17_057	nina_17_057	nina_17_057
42	nina_17_057	nina_17_057	nina_17_057	nina_17_057	nina_17_057	nina_17_057	nina_17_057	nina_17_057	nina_17_057

Sentence 3 - "Young people participate in athletic activities."

	1	2	9	3	5	8	4	6	7
1	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074
2	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074
3	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074
4	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074
5	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074
6	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074
7	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074	nina_11_074
8	nina_17_014	nina_17_014	nina_25_076	nina_05_012	nina_05_012	nina_05_012	nina_05_012	nina_05_012	nina_05_012
9	nina_17_014	nina_17_014	nina_17_014	nina_23_099	nina_23_099	nina_23_099	nina_23_099	nina_23_099	nina_23_099
10	nina_23_055	nina_23_055	nina_23_055	nina_23_099	nina_23_099	nina_23_099	nina_23_099	nina_23_099	nina_23_099
11	nina_23_055	nina_23_055	nina_23_055	nina_23_099	nina_23_099	nina_23_099	nina_23_099	nina_23_099	nina_23_099
12	nina_23_055	nina_23_055	nina_23_055	nina_23_099	nina_23_099	nina_23_099	nina_23_099	nina_23_099	nina_23_099
13	nina_15_040	nina_15_040	nina_15_040	nina_24_032	nina_24_032	nina_24_032	nina_24_032	nina_24_032	nina_24_032
14	nina_21_007	nina_21_007	nina_21_007	nina_21_007	nina_21_007	nina_21_007	nina_21_007	nina_21_007	nina_21_007
15	nina_19_026	nina_19_026	nina_19_026	nina_19_026	nina_19_026	nina_19_026	nina_19_026	nina_19_026	nina_19_026
16	nina_19_026	nina_19_026	nina_19_026	nina_19_026	nina_19_026	nina_19_026	nina_19_026	nina_19_026	nina_19_026
17	nina_19_026	nina_19_026	nina_19_026	nina_19_026	nina_19_026	nina_19_026	nina_19_026	nina_19_026	nina_19_026
18	nina_19_026	nina_19_026	nina_19_026	nina_19_026	nina_19_026	nina_19_026	nina_19_026	nina_19_026	nina_19_026
19	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035
20	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035
21	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035
22	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035
23	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035
24	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035
25	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035	nina_23_035
26	nina_23_035	nina_23_035	nina_23_035	nina_21_039	nina_21_039	nina_21_039	nina_21_039	nina_21_039	nina_21_039
27	nina_24_100	nina_24_100	nina_24_100	nina_07_042	nina_07_042	nina_07_042	nina_07_042	nina_07_042	nina_21_039
28	nina_24_100	nina_24_100	nina_24_100	nina_07_042	nina_07_042	nina_07_042	nina_07_042	nina_07_042	nina_21_039
29	nina_24_100	nina_24_100	nina_24_100	nina_07_042	nina_07_042	nina_07_042	nina_07_042	nina_07_042	nina_07_042
30	nina_24_100	nina_24_100	nina_24_100	nina_07_042	nina_07_042	nina_07_042	nina_07_042	nina_07_042	nina_07_042
31	nina_24_100	nina_24_100	nina_24_100	nina_07_042	nina_07_042	nina_07_042	nina_07_042	nina_07_042	nina_07_042
32	nina_24_100	nina_24_100	nina_24_100	nina_07_042	nina_07_042	nina_07_042	nina_07_042	nina_07_042	nina_07_042
33	nina_24_100	nina_24_100	nina_24_100	nina_07_042	nina_07_042	nina_07_042	nina_07_042	nina_07_042	nina_07_042
34	nina_24_100	nina_24_100	nina_24_100	nina_07_042	nina_07_042	nina_07_042	nina_07_042	nina_07_042	nina_07_042
35	nina_24_100	nina_24_100	nina_24_100	nina_07_042	nina_07_042	nina_07_042	nina_07_042	nina_07_042	nina_07_042

Sentence 4 - "Etiquette mandates compliance with existing regulations"

	1	2	9	3	5	8	4	6	7
1	nina_18_086	nina_18_086	nina_18_086	nina_18_086	nina_18_086	nina_18_086	nina_18_097	nina_18_097	nina_18_097
2	nina_25_070	nina_25_070	nina_25_070	nina_20_038	nina_25_070	nina_25_070	nina_e1_012	nina_e1_012	nina_e1_012
3	nina_25_070	nina_25_070	nina_25_070	nina_20_038	nina_25_070	nina_25_070	nina_e1_012	nina_e1_012	nina_e1_012
4	nina_17_030	nina_17_030	nina_17_030	nina_20_038	nina_25_070	nina_25_070	nina_e1_012	nina_e1_012	nina_e1_012
5	nina_21_007	nina_21_007	nina_21_007	nina_16_067	nina_17_093	nina_17_093	nina_06_006	nina_07_017	nina_07_017
6	nina_25_046	nina_25_046	nina_25_046	nina_05_043	nina_16_064	nina_16_064	nina_05_043	nina_16_082	nina_16_082
7	nina_25_046	nina_25_046	nina_25_046	nina_05_043	nina_01_053	nina_01_053	nina_05_043	nina_05_043	nina_06_035
8	nina_07_038	nina_07_038	nina_07_038	nina_11_079	nina_11_079	nina_11_079	nina_05_043	nina_05_043	nina_11_079
9	nina_07_038	nina_07_038	nina_07_038	nina_06_005	nina_06_005	nina_06_005	nina_14_034	nina_14_034	nina_06_005
10	nina_12_102	nina_12_102	nina_12_102	nina_06_005	nina_06_005	nina_06_005	nina_14_034	nina_14_034	nina_06_005
11	nina_12_102	nina_12_102	nina_12_102	nina_23_047	nina_23_047	nina_23_047	nina_23_047	nina_23_047	nina_23_047
12	nina_08_017	nina_08_017	nina_08_017	nina_20_041	nina_20_041	nina_20_041	nina_20_041	nina_20_041	nina_20_041
13	nina_08_017	nina_08_017	nina_08_017	nina_20_041	nina_20_041	nina_20_041	nina_20_041	nina_20_041	nina_20_041
14	nina_22_046	nina_22_046	nina_22_046	nina_07_068	nina_07_068	nina_07_068	nina_07_068	nina_07_068	nina_07_068
15	nina_13_067	nina_13_067	nina_13_067	nina_07_068	nina_07_068	nina_07_068	nina_07_068	nina_07_068	nina_07_068
16	nina_13_067	nina_13_067	nina_13_067	nina_07_068	nina_07_068	nina_07_068	nina_07_068	nina_07_068	nina_07_068
17	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067
18	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067
19	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067
20	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067
21	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067	nina_13_067
22	nina_03_060	nina_03_060	nina_03_060	nina_03_060	nina_03_060	nina_03_060	nina_03_060	nina_03_060	nina_03_060
23	nina_17_056	nina_17_056	nina_17_056	nina_17_056	nina_17_056	nina_17_056	nina_17_056	nina_17_056	nina_17_056
24	nina_17_056	nina_17_056	nina_17_056	nina_17_056	nina_17_056	nina_17_056	nina_17_056	nina_17_056	nina_17_056
25	nina_18_058	nina_18_058	nina_18_058	nina_11_098	nina_11_098	nina_11_098	nina_11_098	nina_11_098	nina_11_098
26	nina_18_058	nina_18_058	nina_18_058	nina_11_098	nina_11_098	nina_11_098	nina_11_098	nina_11_098	nina_11_098
27	nina_12_086	nina_12_086	nina_12_086	nina_12_087	nina_12_087	nina_12_087	nina_12_087	nina_12_087	nina_12_087
28	nina_12_086	nina_12_086	nina_12_086	nina_12_086	nina_12_086	nina_12_086	nina_12_087	nina_12_087	nina_12_087
29	nina_12_086	nina_12_086	nina_12_086	nina_12_086	nina_12_086	nina_12_086	nina_12_087	nina_12_087	nina_12_087
30	nina_12_086	nina_12_086	nina_12_086	nina_12_086	nina_12_086	nina_12_086	nina_12_087	nina_12_087	nina_12_087
31	nina_12_086	nina_12_086	nina_12_086	nina_12_086	nina_12_086	nina_12_086	nina_12_087	nina_12_087	nina_12_087
32	nina_12_086	nina_12_086	nina_12_086	nina_12_086	nina_12_086	nina_12_086	nina_12_087	nina_12_087	nina_12_087
33	nina_12_086	nina_12_086	nina_12_086	nina_12_086	nina_12_086	nina_12_086	nina_16_097	nina_16_097	nina_16_097
34	nina_17_058	nina_17_058	nina_17_058	nina_17_058	nina_17_058	nina_17_058	nina_16_097	nina_16_097	nina_16_097
35	nina_17_058	nina_17_058	nina_17_058	nina_17_058	nina_17_058	nina_17_058	nina_16_097	nina_16_097	nina_16_097
36	nina_18_027	nina_18_027	nina_18_027	nina_18_027	nina_18_027	nina_18_027	nina_10_044	nina_10_044	nina_10_044
37	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025
38	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025
39	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025
40	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025
41	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025
42	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025
43	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025
44	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025	nina_13_025

Sentence 5 - "Continental drift is a geological theory."

	1	2	9	3	5	8	4	6	7
1	nina_18_054	nina_18_054	nina_18_054	nina_18_054	nina_18_054	nina_18_054	nina_18_054	nina_18_054	nina_18_054
2	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022
3	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022
4	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022
5	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022
6	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022
7	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022
8	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022
9	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022
10	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022	nina_23_022
11	nina_03_018	nina_03_018	nina_03_018	nina_03_018	nina_03_018	nina_03_018	nina_03_018	nina_03_018	nina_03_018
12	nina_18_035	nina_18_035	nina_18_035	nina_18_035	nina_18_035	nina_18_035	nina_18_035	nina_18_035	nina_18_035
13	nina_18_035	nina_18_035	nina_18_035	nina_18_035	nina_18_035	nina_18_035	nina_18_035	nina_18_035	nina_18_035
14	nina_18_035	nina_18_035	nina_18_035	nina_18_035	nina_18_035	nina_18_035	nina_18_035	nina_18_035	nina_18_035
15	nina_18_035	nina_18_035	nina_18_035	nina_18_035	nina_18_035	nina_18_035	nina_18_035	nina_18_035	nina_18_035
16	nina_12_063	nina_12_063	nina_12_063	nina_12_063	nina_12_063	nina_12_063	nina_12_063	nina_12_063	nina_12_063
17	nina_12_063	nina_12_063	nina_12_063	nina_12_063	nina_12_063	nina_12_063	nina_12_063	nina_12_063	nina_12_063
18	nina_04_019	nina_04_019	nina_04_019	nina_04_019	nina_04_019	nina_04_019	nina_04_019	nina_04_019	nina_04_019
19	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080
20	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080
21	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080
22	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080
23	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080	nina_11_080
24	nina_07_081	nina_07_081	nina_11_080	nina_07_081	nina_07_081	nina_19_050	nina_07_081	nina_07_081	nina_11_080
25	nina_07_081	nina_07_081	nina_06_033	nina_07_081	nina_07_081	nina_19_050	nina_07_081	nina_07_081	nina_06_033
26	nina_07_081	nina_07_081	nina_06_033	nina_07_081	nina_07_081	nina_19_050	nina_07_081	nina_07_081	nina_06_033
27	nina_07_081	nina_07_081	nina_06_033	nina_07_081	nina_07_081	nina_19_050	nina_07_081	nina_07_081	nina_06_033
28	nina_r1_058	nina_r1_058	nina_r1_058	nina_r1_058	nina_r1_058	nina_r1_058	nina_r1_058	nina_r1_058	nina_r1_058
29	nina_01_041	nina_01_041	nina_01_041	nina_01_041	nina_01_041	nina_01_041	nina_24_091	nina_24_091	nina_24_091
30	nina_03_088	nina_03_088	nina_03_088	nina_03_088	nina_03_088	nina_03_088	nina_24_091	nina_24_091	nina_24_091
31	nina_04_037	nina_04_037	nina_04_037	nina_04_037	nina_04_037	nina_04_037	nina_24_091	nina_24_091	nina_24_091
32	nina_04_037	nina_04_037	nina_04_037	nina_04_037	nina_04_037	nina_04_037	nina_12_033	nina_12_033	nina_12_033

Bibliography

Adachi, K., Toda, T., Kawanami, H., Saruwatari, H., and Shikano, K. (2005). Designing target cost function based on prosody of speech database. *IEICE Trans. Inf. & Syst.*, E88-D(3).

Allen, P. and Scollie, S. (2002). Stimulus set effects in the similarity ratings of unfamiliar complex sounds. *Journal of the Acoustical Society of America*.

Barry, J., Blamey, P., and Martin, L. (2002). A multidimensional scaling analysis of tone discrimination ability in cantonese-speaking children using a cochlear implant. *Clinical Linguistics & Phonetics*.

Black, A., Taylor, P., and Caley, R. (2000). The festival speech synthesis system, version 2.0. Online. file:///projects/festival/course/doc/festival-html/book1.htm.

Clark, R., Richmond, K., and King, S. (2004). Festival 2 - build your own general purpose unit selection speech synthesiser. *Fifth ISCA Speech Synthesis Workshop*.

Electronic Textbook - StatSoft (2003). Multidimensional scaling. Online. http://www.statsoft.com/textbook/stmulsca.html.

Grabe, E., Rosner, B., Garcia-Albea, J., and Zhou, X. (2003). Perception of english intonation by english, spanish, and chinese listeners. *Language and Speech*.

Hall, J. (2001). Application of multidimensional scaling to subjective evaluation. *Journal of the Acoustical Society of America*.

Hunt, A. and Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. *ICASSP*.

Iverson, P., Kuhl, P., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., and Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*.

King, S., Mayo, C., and Clark, R. (2005). TESS: Testing evaluation of speech synthesis. Not yet published.

Kreiman, J. and Gerratt, B. (1998). Validity of rating scale measures of voice quality. *Journal of the Acoustical Society of America*.

Kruskal, J. and Wish, M. (1978). *Multidimensional Scaling*. Quantitative Applications in the Social Sciences. SAGE Publications Ltd.

Ladd, D. R. (1996). *Intonational Phonology*. Cambridge University Press.

Lee, M., Lopresti, D., and Olive, J. (2001). A text-to-speech platform for variable length optimal unit searching using perceptual cost functions. *4th ISCA Tutorial and Research Workshop on Speech Synthesis*.

Prudon, R. and d'Alessandro, C. (2001). A selection/concatenation text-to-speech synthesis system: databases development, system design, comparative evaluation. *4th ISCA Tutorial and Research Workshop on Speech Synthesis*.

Schneider, W., Eschman, A., and Zuccolotto, A. (2002). *E-Prime User's Guide*. Psychology Software Tools, Inc., University of Pittsburgh.

Stöber, K., Wagner, P., Klabbers, E., and Hess, W. (2001). Definition of a training set for unit selection-based speech synthesis. *4th ISCA Tutorial and Research Workshop on Speech Synthesis*.

Syrdal, A. and Conkie, A. (2004). Data-driven perceptually based join costs. *5th ISCA Speech Synthesis Workshop*.

Vepa, J. and King, S. (2004a). *Join cost for unit selection speech synthesis*, chapter 3. Prentice Hall.

Vepa, J. and King, S. (2004b). Subjective evaluation of join cost and smoothing methods. *5th ISCA Speech Synthesis Workshop*.